Robust inference with incompleteness for logistic regression model

M. Cherifi, M.N. El Korso, Stefano Fortunati, A. Mesloub, L. Ferro-Famil

Robust inference with incompleteness for logistic regression model

M. Cherifi[a], M. N. El Korso[b], S. Fortunati[c], A. Mesloub[a], L. Ferro-Famil[d]

[a] *Lab. Traitement du signal, Ecole Militaire Polytechnique, BP 17 Bordj El Bahri, Algeria*
[b] *University of Paris Saclay, CNRS, CentraleSupelec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France*
[c] *SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, 19 Place Marguerite Perey, 91120, Palaiseau, France*
[d] *ISAE-SUPAERO & CESBIO, Toulouse*

Abstract

Logistic regression models traditionally assume observed covariates. However, practical scenarios often involve missing data and outliers, which pose significant challenges. This short communication presents a new approach to solve these issues by integrating random covariates following a Student t-distribution within the framework of logistic regression. We propose a Robust Stochastic Approximation Expectation-Maximization algorithm suitable for Logistic Regression (REM-LR) that, in addition, is able to improve the resilience of the model against both missing values and outliers.

Keywords: Parametric estimation, missing values, robustness, maximum likelihood, expectation-maximization algorithm, logistic regression

1. Introduction

Logistic regression is a reference statistical method for binary classification, valued for its simplicity and interpretability [1]. It is widely used across various fields: from medicine (to predict diseases) to finance (for credit risk assessment) and social sciences (for behavioral predictions). However, in practice, the efficiency of logistic regression can be compromised by several structural challenges encountered in real-world data, notably: i) the presence of missing data [2, 3] and ii) the presence of outliers in the covariates [4]. Both of these aspects significantly influence parameter estimation and prediction performance.

Missing data is an omnipresent issue during data collection, potentially stemming from various causes: incomplete survey responses, human input errors, or technical failures [5, 6, 7, 8]. Without appropriate handling, these gaps can introduce bias into the estimations, making predictions unreliable and inaccurate. In the literature, managing missing data in logistic regression has often been approached under the assumption that covariates follow a Gaussian distribution. A common method is to maximize the likelihood of missing data under this assumption to estimate the parameters [6]. However, this Gaussian assumption may result inadequate in contexts where data is heterogeneous or contains outliers, which can decrease significantly the estimation accuracy [9, 10]. In contaminated environments, where covariates are affected by outliers, using the Student's t-distribution presents a more robust alternative. Indeed, this distribution, thanks to its heavier tails, offers greater tolerance to extreme values, making it particularly suitable for contaminated data. Moreover, it allows for better modeling of heterogeneous data, a scenario often observed in real-world applications. Thus, in this work, we propose a hybrid approach that combines the modeling of covariates using a heavy-tailed distribution with a robust estimation framework. Specifically, assuming a Student's t-distribution for the covariates enhances the robustness of their parameter estimates. However, this property does not naturally extend to regression parameters, which remain sensitive to outliers. To overcome this

*Corresponding author
Email addresses: cherifi.meddd@gmail.com (M. Cherifi), mohammed.nabil.el-korso@centralesupelec.fr (M. N. El Korso), stefano.fortunati@telecom-sudparis.eu (S. Fortunati), meslouba@gmail.com (A. Mesloub), laurent.ferro-famil@isae-supaero.fr (L. Ferro-Famil)

limitation, we introduce a robust weighting mechanism in the estimation process to mitigate the influence of outliers. More specifically, we adopt the weighting procedure proposed by Carroll and Pederson [11], which adjusts the log-likelihood of the variables of interest to reduce the impact of outlying observations. This strategy ensures that both the covariate distribution and the regression model maintain their robustness in the presence of contaminated data, thereby enhancing the overall reliability of the estimation.

It is important to note that, although the robustness of logistic regression has been studied in several previous works [11, 12, 13], the simultaneous handling of outliers and missing data remains an unresolved challenge. These works primarily focused on robustness against outliers but overlooked the issue of data incompleteness. To fill this gap, we introduce a unified approach that addresses both problems jointly.

In this context, we develop an extension of the stochastic approximation expectation-maximization (SAEM) algorithm, which we call REM-LR (Robust Expectation-Maximization for Logistic Regression). This algorithm is designed to estimate the parameters of logistic regression in the presence of non-Gaussian, contaminated, and potentially missing covariates. REM-LR relies on two key principles: i) robust modeling of covariates using the Student's t-distribution and ii) incorporating robust weighting into the cost function related to variable responses, which minimizes the influence of outliers during parameter estimation.

To evaluate the effectiveness and robustness of our approach, we conducted a series of experiments on synthetic datasets simulating different scenarios: the presence of outliers, missing data under various schemes, and non-Gaussian distribution conditions for the covariates. We compared the performance of REM-LR with that of the stochastic approximation EM and other robust methods that handle missing data. Additionally, we tested our algorithm on real data set. The results demonstrate that REM-LR outperforms existing methods in terms of robustness and accuracy, particularly in scenarios involving contaminated or incomplete data.

2. Model setup

In the following we consider the logistic regression model which reads

$$f_{\boldsymbol{\beta}}(\boldsymbol{x}_i) \triangleq \Pr(y_i = 1 \mid \boldsymbol{x}_i; \boldsymbol{\beta}) = \left(1 + \exp\left(-[1\ \boldsymbol{x}_i^T]\boldsymbol{\beta}\right)\right)^{-1}, \tag{1}$$

for $i = 1, \cdots, N$ where $y_i \in \{0, 1\}$ represents the mutually independent i-th observed binary response variable associated to the i-th covariate vector $\boldsymbol{x}_i \in \mathbb{R}^P$ while $\boldsymbol{\beta} \in \mathbb{R}^{P+1}$ indicates the regression parameters to be estimated. Note that the covariate vectors are assumed to be mutually independent. As experienced in many practical situations, the covariates may contain missing data whose pattern can be characterized by a random mask matrix $\boldsymbol{M}$ such that:

$$[\boldsymbol{M}]_{ip} = \begin{cases} 1 & \text{if } [\boldsymbol{x}_i]_p \text{ is observed,} \\ 0 & \text{if } [\boldsymbol{x}_i]_p \text{ is missing.} \end{cases} \tag{2}$$

In the following, we denote the unobervable full dataset $\mathcal{X} \triangleq \{\mathcal{X}_{\text{obs}}, \mathcal{X}_{\text{mis}}\}$ where $\mathcal{X}_{\text{obs}} \triangleq \{\boldsymbol{x}_{i,\text{obs}}\}_1^N$ is the observed dataset and $\mathcal{X}_{\text{mis}} \triangleq \{\boldsymbol{x}_{i,\text{mis}}\}_1^N$ the missing dataset.
The maximum likelihood maximizes $p(\boldsymbol{y}, \{\boldsymbol{x}_{i,\text{obs}}\}_{i=1}^N, \boldsymbol{M}; \boldsymbol{\psi}) = p(\boldsymbol{y}, \{\boldsymbol{x}_{i,\text{obs}}\}_{i=1}^N; \boldsymbol{\psi}) p(\boldsymbol{M} | \boldsymbol{y}, \{\boldsymbol{x}_{i,\text{obs}}\}_{i=1}^N, ; \boldsymbol{\psi})$ in which $\boldsymbol{\psi}$ represents all the unknown parameters and will be specified in the next section. In this short communication, we limit ourselves to the case of *missing at random (MAR)* and the *missing completely at random (MCAR)* patterns [14]. Along with their incompleteness, we aim at taking into account the possible heavy-tailedness of the (full) covariates and their consequent departure from the classical Gaussian behaviour. Among all the possible heavy-tailed distribution proposed in literature, many studies showed that the Student t-distribution is a good representative able to provide robustness to inference procedures [15, 16]. From the Stochastic Representation Theorem [17], a covariate $\mathcal{X} \ni \boldsymbol{x}_i \sim p(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is t-distributed, with location vector $\boldsymbol{\mu}$, scatter matrix $\boldsymbol{\Sigma}$ and degree of freedom ν controlling the spikiness, *iff*, $\boldsymbol{x}_i | \tau_i \sim p(\boldsymbol{x}_i | \tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}, \tau_i^{-1}\boldsymbol{\Sigma})$, where $\tau_i \sim p(\tau_i; \nu) = \text{Gam}(\nu/2, \nu/2)$ and $\text{Gam}(\cdot, \cdot)$ is the Gamma distribution.

3. Algorithm design

In this short communication, we aim at estimating $\boldsymbol{\beta}$ by relying on the set of response variables $\boldsymbol{y} = [y_1, \cdots, y_N]^T$ and on the *observed* set of covariates $\mathcal{X}_{\text{obs}} \triangleq \{\boldsymbol{x}_{i,\text{obs}}\}_1^N \subseteq \mathcal{X}$, that is a subset of the full set $\mathcal{X} = \{\boldsymbol{x}_i\}_1^N$ of heavy-tailed covariates that are assumed to follow a t-distribution $t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Consequently, the vector of unknown parameters reads $\boldsymbol{\theta} = \left[\boldsymbol{\beta}^T, \boldsymbol{\mu}^T, \text{vech}(\boldsymbol{\Sigma})^T, \nu\right]^T$ while the joint pdf of the response variable y_i and the related full covariate vector is given by $p(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = B(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) \int_0^{+\infty} p(\boldsymbol{x}_i|\tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\tau_i; \nu) d\tau_i$, where the conditional Bernoulli distribution $B(y_i|\boldsymbol{x}_i; \boldsymbol{\beta}) \triangleq f_{\boldsymbol{\beta}}(\boldsymbol{x}_i)^{y_i}[1 - f_{\boldsymbol{\beta}}(\boldsymbol{x}_i)]^{1-y_i}$ and $f_{\boldsymbol{\beta}}(\boldsymbol{x}_i)$, $p(\boldsymbol{x}_i|\tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $p(\tau_i; \nu)$ are defined Section.2. Under the assumption of sample independence, the likelihood of the observed data is given by:

$$
\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{y}, \mathcal{X}_{\text{obs}}, \boldsymbol{M}) = p(\boldsymbol{y}, \{\boldsymbol{x}_{i,\text{obs}}\}_{i=1}^N, \boldsymbol{M}; \boldsymbol{\theta}, \boldsymbol{\phi}) &= \prod_{i=1}^N p(y_i, \boldsymbol{x}_{i,\text{obs}}, \boldsymbol{m}_i; \boldsymbol{\theta}, \boldsymbol{\phi}) \\
&= \prod_{i=1}^N \int p(y_i, \boldsymbol{x}_i, \tau_i, \boldsymbol{m}_i; \boldsymbol{\theta}, \boldsymbol{\phi})\, d\boldsymbol{x}_{i,\text{mis}}\, d\tau_i \\
&= \prod_{i=1}^N \int p(y_i, \boldsymbol{x}_i, \tau_i; \boldsymbol{\theta})\, p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_i; \boldsymbol{\phi})\, d\boldsymbol{x}_{i,\text{mis}}\, d\tau_i. \qquad (3)
\end{aligned}
$$

Considering that the missingness indicator matrix $\boldsymbol{M} \in \{0, 1\}^{N \times P}$ follows a probability mass function given by

$$
p(\boldsymbol{M} \mid \boldsymbol{y}, \mathcal{X}; \boldsymbol{\phi}) = \prod_{i=1}^N p(\boldsymbol{m}_i \mid \boldsymbol{x}_i, y_i; \boldsymbol{\phi}), \qquad (4)
$$

Here, $\boldsymbol{\phi}$ denotes an unknown deterministic nuisance parameter vector that characterizes the missing data mechanism. Under the MCAR assumption, missingness is independent of both observed and unobserved data, implying $p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_i; \boldsymbol{\phi}) = p(\boldsymbol{m}_i \mid \boldsymbol{\phi})$. In contrast, under the MAR assumption, the missingness pattern depends solely on the observed values $\boldsymbol{x}_{i,\text{obs}}$ and y_i, leading to $p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_i, ; \boldsymbol{\phi}) = p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\phi})$. In the MAR case, the maximization of (3) can be expressed as:

$$
\begin{aligned}
(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}) &= \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \prod_{i=1}^N \int p(y_i, \boldsymbol{x}_i, \tau_i; \boldsymbol{\theta})\, p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\phi})\, d\boldsymbol{x}_{i,\text{mis}} d\tau_i \\
&= \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \prod_{i=1}^N \left[p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\phi}) \int p(y_i, \boldsymbol{x}_i, \tau_i; \boldsymbol{\theta})\, d\boldsymbol{x}_{i,\text{mis}} d\tau_i \right] \\
&= \arg\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \sum_{i=1}^N \left[\log p(\boldsymbol{m}_i \mid y_i, \boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\phi}) + \log \int p(y_i, \boldsymbol{x}_i, \tau_i; \boldsymbol{\theta})\, d\boldsymbol{x}_{i,\text{mis}} d\tau_i \right]. \qquad (5)
\end{aligned}
$$

This result holds true in the MCAR case as well. Consequently, the estimation of $\boldsymbol{\theta}$ can be carried out without the need for explicit modeling of the missing data mechanism, thereby simplifying the process and leading to:

$$
\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^N \log p(y_i, \boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\theta}). \qquad (6)
$$

In contrast, in the MNAR (Missing Not At Random) case, the distribution of the missing data mask, $p(\boldsymbol{M} \mid \boldsymbol{y}, \mathcal{X}; \boldsymbol{\phi})$, must be explicitly modeled, for instance using logistic regression or any suitable model. This introduces a non-separable problem, requiring the estimation of additional parameters to capture the missing data mechanism, which significantly complicates the estimation process. Addressing this challenge

represents a promising avenue for future research. Consequently, under the i.i.d. assumption and assuming either MAR or MCAR missing data patterns, the log-likelihood function for $\boldsymbol{\theta}$ can be expressed as:

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathcal{X}_{\text{obs}}) = \sum_{i=1}^{N} \ln \left(\int B(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) \int p(\boldsymbol{x}_i \mid \tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\tau_i; \nu) d\tau_i \, d\boldsymbol{x}_{i,\text{mis}} \right). \tag{7}$$

Unfortunately, a directly maximization of $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}, \mathcal{X}_{\text{obs}})$ is not feasible due to the impossibility to derive a closed form solution for the marginal density of the observed data.

3.1. Robustness of covariates parameters estimation

In order to handle t-distributed data with missing observations, one possible approach generally used in literature is the Expectation-Maximization (EM) algorithm. We define the *complete* data as $\{\boldsymbol{y}, \mathcal{X}, \boldsymbol{\tau}\}$ such that the related *complete* log-likelihood function is :

$$\begin{aligned} \mathcal{L}_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathcal{X}, \boldsymbol{\tau}) &= \sum_{i=1}^{N} \ln p(y_i, \boldsymbol{x}_i, \tau_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{N} \ln \left[B(y_i \mid \boldsymbol{x}_i; \boldsymbol{\beta}) p(\boldsymbol{x}_i \mid \tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\tau_i; \nu) \right], \end{aligned} \tag{8}$$

with $\boldsymbol{\tau} = [\tau_1, \dots \tau_N]^T$. One can immediately note that the *complete* joint density $p(y_i, \boldsymbol{x}_i, \tau_i; \boldsymbol{\theta})$ belongs to the exponential family of distributions [18] since it is the product of the (conditional) Bernoulli $B(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})$, the (conditional) Gaussian $p(\boldsymbol{x}_i|\tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the Gamma distributions $p(\tau_i; \nu)$ that are exponential distribution themselves allowing us to work directly with the complete-data sufficient statistics [19].

The E-step of the EM algorithm consists in computing the following surrogate at the $t+1$-th iteration

$$Q^{(t+1)}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\tau}, \mathcal{X}_{\text{mis}} | \boldsymbol{y}, \mathcal{X}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(t)}} \left\{ \mathcal{L}_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathcal{X}, \boldsymbol{\tau}) \right\} \tag{9}$$

$$= \mathbb{E}_{\mathcal{X}_{\text{mis}} | \boldsymbol{y}, \mathcal{X}_{\text{obs}}} \left\{ g^{(t+1)}(\boldsymbol{y}, \mathcal{X}, \boldsymbol{\theta}) \right\} \text{ where:} \tag{10}$$

$$g^{(t+1)}(\boldsymbol{y}, \mathcal{X}, \boldsymbol{\theta}) \triangleq \mathbb{E}_{\boldsymbol{\tau} | \boldsymbol{y}, \mathcal{X}, \hat{\boldsymbol{\theta}}^{(t)}} \left\{ \mathcal{L}_c(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathcal{X}, \boldsymbol{\tau}) \right\} \tag{11}$$

Evaluating in close form the expectation in (9) is impossible and even its numerical calculation may be problematic since it require the generation of random samples from the joint distribution of $(\boldsymbol{\tau}, \mathcal{X}_{\text{mis}})$. One may rely then on the Stochastic Approximation EM (SAEM) algorithm, introduced in [19], as a fast and computationally efficient version of the classical EM methods. The SAEM algorithm is structured in three steps: the *simulation*, *stochastic approximation* and the *maximization* steps.

The simulation step: In the expectation (11), we substitute the unavailable set of full covariates $\mathcal{X}$ with a "simulated" version as:

$$\tilde{g}^{(t+1)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\tau} | \boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}} \left\{ \mathcal{L}_c(\boldsymbol{\theta} | \boldsymbol{y}, \mathcal{X}, \boldsymbol{\tau}) \right\} \tag{12}$$

where $\widehat{\mathcal{X}}^{(t)} \triangleq \{\hat{\boldsymbol{x}}_i^{(t)}\}_{i=1}^{N}$ in which:

$$[\hat{\boldsymbol{x}}_i^{(t)}]_p = \begin{cases} [\boldsymbol{x}_{i,\text{obs}}]_p & [\boldsymbol{M}]_{ip} = 1, \\ [\hat{\boldsymbol{x}}_{i,\text{mis}}^{(t)} \sim p(\boldsymbol{x}_{i,\text{mis}} | \boldsymbol{y}, \mathcal{X}_{\text{obs}}, \hat{\boldsymbol{\theta}}^{(t)})]_p & [\boldsymbol{M}]_{ip} = 0. \end{cases}$$

It is worth noticing that the samples $\hat{\boldsymbol{x}}_{i,\text{mis}}^{(t)}$ can be obtained by using a Metropolis-Hastings procedure, as described in the supporting materials.

By substituting (8) in (12), we get:

$$\tilde{g}^{(t+1)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}, \boldsymbol{\theta}) = \tilde{S}_{\boldsymbol{\beta}}^{(t+1)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) + \mathbb{E}_{\boldsymbol{\tau} | \boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}} \left\{ \ln \left[p(\hat{\boldsymbol{x}}_i^{(t)} | \tau_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\tau_i; \nu) \right] \right\}, \tag{13}$$

$$\tilde{S}_{\boldsymbol{\beta}}^{(t+1)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) \triangleq \sum\nolimits_{i=1}^{N} \ln B(y_i|\hat{\boldsymbol{x}}_i^{(t)}; \boldsymbol{\beta}). \tag{14}$$

First, note that the parameters of interest $\boldsymbol{\beta}$ and the t-distribution parameters ($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and ν), are decoupled, then they can be handled separately. Second, since the complete-data joint pdf belongs to the exponential family, instead of working directly with the function $\tilde{g}^{(t+1)}$, we may use the complete-data sufficient statistics [19]. Specifically, following [15], complete-data sufficient statistics are

$$S_{\boldsymbol{\tau x}}^{(t)}(\widehat{\mathcal{X}}^{(t)}) = \sum\nolimits_{i=1}^{N} \tau_i \hat{\boldsymbol{x}}_i^{(t)},\ S_{\boldsymbol{\tau\tau}}^{(t)} = \sum\nolimits_{i=1}^{N} (\ln(\tau_i) - \tau_i)$$

$$S_{\boldsymbol{\tau}}^{(t)} = \sum\nolimits_{i=1}^{N} \tau_i,\ S_{\boldsymbol{\tau xx}}^{(t)}(\widehat{\mathcal{X}}^{(t)}) = \sum\nolimits_{i=1}^{N} \tau_i \hat{\boldsymbol{x}}_i^{(t)} [\hat{\boldsymbol{x}}_i^{(t)}]^T,$$

According to (13), we need to update these sufficient statistics by evaluating their expectation $\mathbb{E}_{\boldsymbol{\tau}|\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}}\{\cdot\}$ which read

$$\tilde{S}_{\boldsymbol{\tau}}^{(t+1)}(\widehat{\mathcal{X}}^{(t)}) = \sum\nolimits_{i=1}^{N} w_i^{(t+1)}, \tag{15}$$

$$\tilde{S}_{\boldsymbol{\tau x}}^{(t+1)}(\widehat{\mathcal{X}}^{(t)}) = \sum_{i=1}^{N} w_i^{(t+1)} \hat{\boldsymbol{x}}_i^{(t)}, \tag{16}$$

$$\tilde{S}_{\boldsymbol{\tau xx}}^{(t+1)}(\widehat{\mathcal{X}}^{(t)}) = \sum\nolimits_{i=1}^{N} w_i^{(t+1)} \hat{\boldsymbol{x}}_i^{(t)} [\hat{\boldsymbol{x}}_i^{(t)}]^T, \tag{17}$$

$$\tilde{S}_{\boldsymbol{\tau\tau}}^{(t+1)}(\widehat{\mathcal{X}}^{(t)}) = N[\phi((P + \hat{\nu}^{(t)})/2) - \ln((P + \hat{\nu}^{(t)})/2)] + \sum\nolimits_{i=1}^{N} [\ln(w_i^{(t+1)}) - w_i^{(t+1)}]. \tag{18}$$

where $\phi(t) = \frac{d}{dt} \ln \Gamma(t)$ and $w_i^{(t+1)} = \mathbb{E}_{\boldsymbol{\tau}|\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}, \hat{\boldsymbol{\theta}}^{(t)}}\{\tau_i\} = (\hat{\nu}^{(t)} + P)/(\hat{\nu}^{(t)} + \hat{\delta}_i^{(t)})$, where $\hat{\delta}_i^{(t)}$ indicated the squared Mahalanobis distance $\hat{\delta}_i^{(t)} = (\hat{\boldsymbol{x}}_i^{(t)} - \hat{\boldsymbol{\mu}}^{(t)})^T [\hat{\boldsymbol{\Sigma}}^{(t)}]^{-1} (\hat{\boldsymbol{x}}_i^{(t)} - \hat{\boldsymbol{\mu}}^{(t)})$.

The stocastic approximation step: In order to avoid the analytical difficulties in evaluating the conditional expectation in (9) and since, the joint complete data pdf belongs to the exponential family, instead of maximizing the objective function $Q^{(t+1)}(\boldsymbol{\theta})$, we can maximize the data sufficient statistics obtained from the following stochastic approximations:

$$\hat{S}_{\boldsymbol{\tau}}^{(t+1)} = \hat{S}_{\boldsymbol{\tau}}^{(t)} + \alpha_t \left(\tilde{S}_{\boldsymbol{\tau}}^{(t+1)} - \hat{S}_{\boldsymbol{\tau}}^{(t)} \right), \tag{19}$$

$$\hat{S}_{\boldsymbol{\tau\tau}}^{(t+1)} = \hat{S}_{\boldsymbol{\tau\tau}}^{(t)} + \alpha_t \left(\tilde{S}_{\boldsymbol{\tau\tau}}^{(t+1)} - \hat{S}_{\boldsymbol{\tau\tau}}^{(t)} \right), \tag{20}$$

$$\hat{S}_{\boldsymbol{\tau x}}^{(t+1)} = \hat{S}_{\boldsymbol{\tau x}}^{(t)} + \alpha_t \left(\tilde{S}_{\boldsymbol{\tau x}}^{(t+1)} - \hat{S}_{\boldsymbol{\tau x}}^{(t)} \right), \tag{21}$$

$$\hat{S}_{\boldsymbol{\tau xx}}^{(t+1)} = \hat{S}_{\boldsymbol{\tau xx}}^{(t)} + \alpha_t \left(\tilde{S}_{\boldsymbol{\tau xx}}^{(t+1)} - \hat{S}_{\boldsymbol{\tau xx}}^{(t)} \right). \tag{22}$$

where, for notation simplicity, the dependence of all the previous terms from $\widehat{\mathcal{X}}^{(t)}$ has been omitted. According to the same sthocastic approximation principle, we have:

$$\hat{S}_{\boldsymbol{\beta}}^{(t+1)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) = \hat{S}_{\boldsymbol{\beta}}^{(t)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) + \alpha_t \left(\tilde{S}_{\boldsymbol{\beta}}^{(t+1)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) - \hat{S}_{\boldsymbol{\beta}}^{(t)}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) \right) \tag{23}$$

We note in passing that the choice of the step size α_t is critical for the convergence of the algorithm. Here we choose to initially set $\{\alpha_t\}$ to 1 in the first iterations, followed by a decreasing sequence such as $1/t$.

The maximization step: As explained above, in our case the SAEM algorithm allows for a derivation of $\hat{\boldsymbol{\theta}}^{(t+1)}$ from the stochastic approximation of the sufficient statistics. Specifically, an estimation the t-distribution parameter can be obtained from eqs. (19)-(22) as:

$$\hat{\boldsymbol{\mu}}^{(t+1)} = \frac{\hat{S}_{\boldsymbol{\tau x}}^{(t+1)}}{\hat{S}_{\boldsymbol{\tau}}^{(t+1)}}, \tag{24}$$

$$\hat{\boldsymbol{\Sigma}}^{(t+1)} = \frac{1}{N} \left(\hat{S}_{\boldsymbol{\tau xx}}^{(t+1)} - \frac{\hat{S}_{\boldsymbol{\tau x}}^{(t+1)} (\hat{S}_{\boldsymbol{\tau x}}^{(t+1)})^T}{\hat{S}_{\boldsymbol{\tau}}^{(t+1)}} \right), \tag{25}$$

while the degree of freedom estimation requires solving the following equation using the quasi-Newton algorithm:

$$-\phi(\nu/2) + \ln(\nu/2) + 1 + N^{-1}\hat{S}^{(t+1)}_{\boldsymbol{\tau}\boldsymbol{\tau}}\Big|_{\nu=\nu^{(t+1)}} = 0, \tag{26}$$

the regressor parameter estimates are obtained by solving:

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \arg\max_{\boldsymbol{\beta}} \hat{S}^{(t+1)}_{\boldsymbol{\beta}}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}), \tag{27}$$

which is also solved using quasi-Newton algorithm. The *REM-LR* algorithm is structured analogously to *SAEM*. The sequence of step sizes $\{\alpha_t\}$ in *REM-LR* satisfies the classical stochastic approximation conditions, namely $\sum_{t=1}^{\infty} \alpha_t = +\infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < +\infty$, which, according to Delyon et al. (1999), ensure its almost sure convergence. The algorithm stops when the squared distance between the estimated parameters $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ at iterations t and $t-10$ falls below a predefined convergence threshold of 10^{-8}. Additionally, a maximum number of iterations is set to $T = 200$ to prevent infinite loops in case of slow convergence.

3.2. Robustness of regressor parameters estimation

The assumption of a Student t-distribution for the covariates leads to robust (in the M-sense) estimators for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ [17], given in eqs. (24) and (25). However, this property does not apply to the estimation of the parameters of interest $\boldsymbol{\beta}$ since the related sufficient statistic in (14) is still vulnerable to outliers. To avoid this problem, a standard procedure in robust statistics is to adopt a weighting function to penalize terms produced by outliers. Specifically, $\tilde{S}^{(t+1)}_{\boldsymbol{\beta}}$ in (14) may be robustified as:

$$\tilde{R}^{(t+1)}_{\boldsymbol{\beta}}(\boldsymbol{y}, \widehat{\mathcal{X}}^{(t)}) \triangleq \sum\nolimits_{i=1}^{N} W(\hat{\delta}_i) \ln B(y_i|\hat{\boldsymbol{x}}_i^{(t)}; \boldsymbol{\beta}), \tag{28}$$

where the decreasing function $W(u)$ must be bounded [20]. The aim of this weighting function is to penalize terms generated by $\hat{\boldsymbol{x}}_i^{(t)}$ that are far from the estimated mean $\hat{\boldsymbol{\mu}}^{(t)}$ in the Mahalanobis distance. Carroll and Pederson [11] proposed selecting W from a specific family determined by a parameter $c > 0$ as $W(u) = \left(1 - \frac{u^2}{c^2}\right)^3 \mathbb{I}(|u| \leq c)$ where $\mathbb{I}$ denotes the indicator function, equal to one if the condition $|u| \leq c$ holds and zero otherwise. In real-world scenarios, data types and outlier characteristics can vary, making it challenging to choose the optimal threshold for the parameter c in the robust weighting function. To address this, we suggest using cross-validation with a K-fold method across a range of possible values for c to determine the best adjustment [21].

3.3. Prediction with missing data

It is common to assume that the distribution of the observed data and the distribution of the missing data are identical to that of the training data. It is than natural to employ a Monte Carlo approach for the prediction step. Namely, by exploiting S Monte Carlo samples using a Metropolis-Hastings (MH) procedure (c.f. subsection 3.4 below), we can directly estimate the response by maximizing its marginal distribution over the missing data, while considering the observed data:

$$\begin{aligned}
\hat{y}_i &= \arg\max_{y_i} B(y_i|\boldsymbol{x}_{i,\text{obs}}; \hat{\boldsymbol{\beta}}) \\
&= \arg\max_{y_i} \int B(y_i|\boldsymbol{x}_{i,\text{obs}}, \hat{\boldsymbol{x}}^{(s)}_{i,\text{mis}}; \hat{\boldsymbol{\beta}}) p(\hat{\boldsymbol{x}}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}; \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}, \hat{\nu}) d\hat{\boldsymbol{x}}_{i,\text{mis}} \\
&= \arg\max_{y_i} \mathbb{E}_{\hat{\boldsymbol{x}}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}; \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}}, \hat{\nu}} \left\{ B(y_i|\boldsymbol{x}_{i,\text{obs}}, \hat{\boldsymbol{x}}^{(s)}_{i,\text{mis}}; \hat{\boldsymbol{\beta}}) \right\} \\
&= \arg\max_{y_i} \sum_{s=1}^{S} B(y_i|\boldsymbol{x}_{i,\text{obs}}, \hat{\boldsymbol{x}}^{(s)}_{i,\text{mis}}; \hat{\boldsymbol{\beta}}).
\end{aligned} \tag{29}$$

3.4. Sampling with the Metropolis-Hastings Algorithm

In the logistic regression scenario, exact sampling of unobserved data from the conditional distribution $p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}, y_i; \boldsymbol{\theta})$ is generally impractical due to its dependence on an integral that does not admit a closed-form solution. A potential solution involves the use of a MH algorithm, which entails constructing a Markov chain with the target distribution as its stationary distribution. The states of the chain, obtained after S iterations, can then be used as a sample from the target distribution. To define a proposal distribution for the MH algorithm, it is observed that the target distribution $p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}, y_i; \boldsymbol{\theta}) \propto B(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. We choose the proposal distribution for $\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}$ as the second term, characterized by a Student's t-distribution $\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}} \sim t_{p_{\text{i,mis}}}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i)$, where $p_{\text{i,mis}} + p_{\text{i,obs}} = p$, the mean vector is $\boldsymbol{\mu}_p = [\boldsymbol{\mu}_{\text{i,mis}}, \boldsymbol{\mu}_{\text{i,obs}}]$, and the scale matrix is

$$\boldsymbol{\Sigma}_{p\times p} = \begin{bmatrix} \boldsymbol{\Sigma}_{\text{i,obs,obs}} & \boldsymbol{\Sigma}_{\text{i,obs,mis}} \\ \boldsymbol{\Sigma}_{\text{i,mis,obs}} & \boldsymbol{\Sigma}_{\text{i,mis,mis}} \end{bmatrix}.$$

Then, we have:

$$\boldsymbol{x}_{\text{i,mis}}|\boldsymbol{x}_{\text{i,obs}} \sim t_{p_{\text{i,mis}}}\left(\boldsymbol{\mu}_{i,\text{mis|obs}}, \frac{\nu + \delta^2_{\text{i,obs}}}{\nu + p_{\text{i,obs}}}\boldsymbol{\Sigma}_{i,\text{mis,mis|obs}}, \nu + p_{\text{i,obs}}\right), \tag{30}$$

$$\boldsymbol{\mu}_{i,\text{mis|obs}} = \boldsymbol{\mu}_{i,\text{mis}} + \boldsymbol{\Sigma}_{i,\text{mis,obs}}\boldsymbol{\Sigma}^{-1}_{i,\text{obs,obs}}(\boldsymbol{x}_{\text{i,obs}} - \boldsymbol{\mu}_{i,\text{obs}}), \tag{31}$$

$$\boldsymbol{\Sigma}_{i,\text{mis,mis|obs}} = \boldsymbol{\Sigma}_{i,\text{mis,mis}} - \boldsymbol{\Sigma}_{i,\text{mis,obs}}\boldsymbol{\Sigma}^{-1}_{i,\text{obs,obs}}\boldsymbol{\Sigma}_{i,\text{obs,mis}}, \tag{32}$$

$$\boldsymbol{\Sigma}_i = \frac{\nu + \delta^2_{\text{i,obs}}}{\nu + p_{\text{i,obs}}}\boldsymbol{\Sigma}_{i,\text{mis,mis|obs}}, \nu_i = \nu + p_{\text{i,obs}}, \tag{33}$$

$$\delta^2_{\text{i,obs}} = (\boldsymbol{x}_{\text{i,obs}} - \boldsymbol{\mu}_{i,\text{obs}})^T\boldsymbol{\Sigma}^{-1}_{i,\text{obs,obs}}(\boldsymbol{x}_{\text{i,obs}} - \boldsymbol{\mu}_{i,\text{obs}}), \tag{34}$$

where $\delta^2_{\text{i,obs}}$ is the squared Mahalanobis distance of $\boldsymbol{x}_{\text{i,obs}}$ from $\boldsymbol{\mu}_{i,\text{obs}}$ with scale matrix $\boldsymbol{\Sigma}_{i,\text{obs,obs}}$.

Knowing that the proposal distribution follows a Student's t-distribution conditioned on the observed variables $\boldsymbol{x}_{\text{i,obs}}$, the proposal distribution is formulated as follows:

$$h(\boldsymbol{x}_{i,\text{mis}}) = p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = t_{p_{\text{i,mis}}}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i).$$

Noting that

$$g(\boldsymbol{x}_{i,\text{mis}}) = p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) \text{ with fixed } \nu \quad \text{and} \quad f(\boldsymbol{x}^{(s)}_{i,\text{mis}}) = p(\boldsymbol{x}_{i,\text{mis}}|\boldsymbol{x}_{i,\text{obs}}, y_i; \boldsymbol{\theta}),$$

we have than all the ingredients to use the MH sampling.

3.5. Complexity of the REM-LR Algorithm

The dominant component of the *REM-LR* algorithm is the Metropolis-Hastings sampling, whose complexity is given by:

$$O(M\text{-}H) \equiv O\left(\sum_{i\in I(\mathcal{X}_{\text{miss}})}\left(p_{\text{obs}}(i)^3 + p_{\text{obs}}(i)\,p_{\text{miss}}(i)^2 + p_{\text{miss}}(i)^3 + B\,p_{\text{miss}}(i)^3\right)\right)$$

where $I(\mathcal{X}_{\text{miss}})$ denotes the set of samples with at least one missing variable and $n_{\text{mis}} = |I(\mathcal{X}_{\text{miss}})|$ is their number, The constant B represents the burn-in period in the Metropolis-Hastings (M-H) sampling process. It is the number of initial iterations discarded to allow the algorithm to converge to the target distribution. For each sample i, since $p_{\text{obs}}(i) = p - p_{\text{miss}}(i)$, substituting this into the complexity expression yields

$$O(M\text{-}H) \equiv O\left(p^3\,n_{\text{mis}} - 3p^2\sum_{i\in I(\mathcal{X}_{\text{miss}})}p_{\text{miss}}(i) + 4p\sum_{i\in I(\mathcal{X}_{\text{miss}})}p_{\text{miss}}(i)^2 + (B-1)\sum_{i\in I(\mathcal{X}_{\text{miss}})}p_{\text{miss}}(i)^3\right).$$

In addition to the *Metropolis-Hastings* phase, the *REM-LR* algorithm comprises an E-step with a computational cost of $O(np^3)$, primarily due to the calculation of the Mahalanobis distance for robust weighting functions, which requires inverting $p \times p$ matrices for each of the n samples. The M-step, on the other hand, has a lower complexity of $O(np^2)$, as it is mainly associated with updating the model parameters. Consequently, over T_{Stop} iterations, the total computational complexity is $O\big(T_{\text{Stop}}(np^2 + np^3 + p^3\, n_{\text{mis}} - 3p^2 \sum_{i\in I(\mathcal{X}_{\text{miss}})} p_{\text{miss}}(i) + 4p \sum_{i\in I(\mathcal{X}_{\text{miss}})} p_{\text{miss}}(i)^2 + (B-1) \sum_{i\in I(\mathcal{X}_{\text{miss}})} p_{\text{miss}}(i)^3)\big)$. In the worst-case scenario, where most variables are missing in each sample (i.e., $p_{\text{miss}}(i) \sim p$ for $i \in I(\mathcal{X}_{\text{miss}})$) and assuming $n_{\text{mis}} \leq n$, the dominant terms remain of order p^3. In summary, while the algorithm is polynomial in p (specifically cubic in p) and linear in n, it can quickly become computationally expensive for large p. Optimizing costly operations—such as matrix inversion—is therefore crucial for efficient performance in high-dimensional applications.

4. Numerical Simulations

4.1. Generation of Synthetic Dataset

In this subsection, our primary focus lies in testing our algorithm in the case of covariates generated from a Gaussian distribution with outliers and missing covariates [22].The regressor parameters $\boldsymbol{\beta}$ and the mean vector $\boldsymbol{\mu}$ is given by $\boldsymbol{\beta} = [-0.3\ 0.4\ \ -0.2\ 0.9\ 0\ \ -0.5]^T$ and $\boldsymbol{\mu} = [1\ 2\ 3\ 4\ 5]^T$, while the covariance matrix is

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1.2 & 0 & 0 & 0 \\ 1.2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 9 & 6 & 12 \\ 0 & 0 & 6 & 16 & 14 \\ 0 & 0 & 12 & 14 & 25 \end{pmatrix}.$$

The *outliers* are generated from multivariate Gaussian distributions with an identity covariance matrix and centers shifted by ± 30 times the covariates means $\boldsymbol{\mu}$.

Table 1: Bias and RMSE for 10% MAR and 10% of outliers

	REM-LR	SAEM	R-MICE	R-Mean	P-M	BY-M	RSI	R-Full
bias(β_1)	-0.0094	-0.1588	0.0186	0.0282	-0.2405	-0.1265	0.0282	-0.0060
bias(β_2)	-0.0216	0.2186	-0.0180	0.0045	0.1971	0.2557	0.0127	0.0026
bias(β_3)	0.0021	-0.0028	0.0102	0.0116	-0.0387	0.0147	0.0126	-0.0061
bias(β_4)	0.0166	-0.0668	-0.0105	-0.0756	0.0472	-0.1832	-0.0731	0.0073
bias(β_5)	-0.0016	0.0046	-0.0008	-0.0124	-0.0046	-0.0091	-0.0092	-0.0026
bias(β_6)	-0.0080	0.0413	0.0063	0.0510	-0.0138	0.1162	0.0456	-0.0017
RMSE($\boldsymbol{\beta}$)	**0.0809**	0.1007	0.0826	0.1177	0.1277	0.1756	0.1138	0.0600

4.2. Hyperparameter Selection

Smoothing Parameter Selection (α_t). For the selection of the smoothing parameter α_t in *REM-LR*, we adopted the sequence $\alpha_t = (t_1 - t)^{-\tau}$ proposed by Kuhn and Lavielle (2004) in the *SAEM* algorithm. During the first 50 iterations ($t_1 = 50$), α_t was set to 1 to accelerate convergence toward the maximum likelihood estimator (MLE). Subsequently, it followed a decreasing schedule controlled by τ, ensuring convergence. We chose $\tau = 1$, an optimal trade-off between speed and stability, which is often recommended in the literature.

Threshold Adjustment (c). The adjustment of the threshold c in the weighting function was performed using 5-fold cross-validation. This process aimed to identify the optimal value of c that maximized predictive performance of the *REM-LR* in the training set. By systematically evaluating different values of c across the folds, we determined that the best predictive performance with this type of data was achieved at $c = 21$. For both simulations, the dataset was corrupted with 10% outliers and 10% MAR missing data.

4.3. Parameter estimation

We present estimates of the biases for β_i as well as the Root Mean Squared Error (RMSE) for $\boldsymbol{\beta}$ across different scenarios. The proposed *REM-LR* is compared with the classical *SAEM* algorithm [6], the robust version of *R-MICE* (for Multiple Imputation by Chained Equations [23] coupled with a weighted logistic regression (28)), the multiple robust stochastic imputation *RSI*[1], the *P-M* Pregibon [25] and Bianco & Yohai *BY-M* [13] methods which are known to be robust in the context of logistic regression models (working without missing data) in which we coupled them with a robust mean imputation of missing data using the MCD method [24]. Finally, we added as benchmark the *R-Full* for the robust full case (i.e., the missing values are considered known).

Table 1 shows that the proposed *REM-LR* exhibits the smallest bias and the lowest RMSE for the regression parameters in the case of partially missing Gaussian covariates corrupted by outliers. Extensive numerical simulations show that the same behavior is observed for different missing data mechanisms (i.e., MCAR, MAR) and for varying percentages of missing values and/or outliers, where the results are presented in Table (4).

To assess our method under an MNAR scenario, we implemented a self-masking mechanism across all features. In this approach, for each feature, any value that exceeds the feature's mean is designated as missing, ensuring that the probability of missingness is inherently dependent on the value itself, a hallmark of a non-ignorable missing data process. We then conducted simulations using MNAR (self-masking) data with a 30% missing rate and 10% outliers. The outcomes reveal that even under MNAR conditions, our algorithm, originally crafted for ignorable missing mechanisms (MCAR and MAR), achieves a lower overall RMSE for $\boldsymbol{\beta}$ compared to other methods (see Figure 1).

4.4. Prediction with Corrupted & Missing Values

To evaluate the predictive accuracy on a test dataset containing missing values, we used the same simulation configurations as detailed in the previous subsection, employing a training set with dimensions 1000×5. Additionally, a test set with dimensions 250×5 was generated. In all imputation scenarios, the missing data were imputed independently for the test set, and then the selected model was applied from the training set. We conducted a comprehensive comparison of these methodologies using conventional criteria to evaluate the quality of predicted probabilities derived from logistic regression. The evaluation metric used was **accuracy**. The average performance indicates a notable superiority of the proposed *REM-LR* algorithm, even with an increase in missing data, as shown in Fig. 2.

4.5. Scalability and Computational Performance

To assess the scalability and computational efficiency of *REM-LR*, we conducted experiments by varying both the sample size (n) and the number of covariates (p). The goal is to evaluate how *REM-LR* performs in terms of execution time, accuracy, and RMSE compared to *MCEM(Monte Carlo Expectation Maximization), SAEM, and BY-M.*

Impact of Sample Size. Table 2 presents a comparative analysis of different estimators for varying sample sizes (n). The results show that the execution time of *REM-LR* scales approximately linearly with n, indicating good scalability for medium-sized datasets. Moreover, *REM-LR* achieves lower RMSE and higher accuracy than *MCEM* and *BY-M*, while maintaining a computational cost close to *SAEM.*

Impact of Number of Covariates. To further evaluate the scalability of *REM-LR*, we measured execution time while varying the number of covariates (p) for a fixed sample size ($n = 1000$). Table 3 illustrates that computation time increases cubically with p, reflecting the growing complexity of the model as the number of covariates increases. However, *REM-LR* remains competitive with *SAEM* in terms of execution time, while providing more accurate estimates.

[1]Robust in the sens of robust imputation, i.e, the mean and variance are robustly estimated using the Minimum Covariance Determinant (MCD) method [24], and then, the imputation are given as a sampling from a Gaussian distribution with a robust mean and covariance matrix estimation using MCD.

Table 2: Comparison of Mean RMSE, Accuracy, and Execution Time for Different Estimators ($p = 5$, 5 simulations) under 10% MCAR Missing Data and 20% Outliers.

Sample Size	RMSE				Accuracy				Execution Time (seconds)			
	BY-M	MCEM	SAEM	REM-LR	BY-M	MCEM	SAEM	REM-LR	BY-M	MCEM	SAEM	REM-LR
200	0.3618	0.4120	0.4066	0.4786	0.6894	0.6900	0.6880	0.6200	16.16	209.99	8.86	9.26
500	0.3664	0.3102	0.3238	0.2416	0.6840	0.6936	0.6928	0.6800	42.82	445.66	20.63	22.28
700	0.3599	0.3395	0.3417	0.1432	0.6667	0.6857	0.6914	0.6938	64.61	623.31	30.16	32.26
1000	0.3587	0.2801	0.3644	0.1624	0.7280	0.6952	0.7105	0.7267	100.81	1089.07	84.67	82.46
3000	0.3565	0.3692	0.3690	0.1301	0.7080	0.7000	0.6898	0.7256	298.72	4664.98	216.07	257.94
5000	0.3383	0.2550	0.3545	0.1039	0.7112	0.7111	0.7045	0.7320	491.99	6180.22	453.25	476.36

Table 3: Comparison of Mean Execution Time (in seconds) for Different Estimators ($n = 1000$, 5 simulations) under 10% MCAR Missing Data and 20% Outliers.

Covariates (p)	MCEM	SAEM	REM-LR
10	6207.6	190.02	202.19
15	7137.8	263.55	287.23
20	8562.4	314.68	338.90
25	11211.0	410.11	428.54
30	15907.0	480.74	463.82
50	20476.1	646.75	634.45

4.6. Application on real dataset: Raisin Dataset

In the scope of our study, we utilized data sourced from computer vision research aimed at distinguishing between two varieties of dried grapes, namely Kecimen and Besni, cultivated in Turkey. The dataset employed consisted of a total of 900 grape grains, evenly distributed between these two varieties after preprocessing [26]. To evaluate and demonstrate the robustness and performance of our algorithm, we employed various machine learning models including Support Vector Machine (*SVM*), *K-Nearest Neighbors (KNN)*, and Random Forest (*RF*). These models were trained and tested on data imputed using the *MissForest* method [27]. The hyperparameters of these machine learning algorithms were adjusted using the cross-validation method with 5 folds in the training phase. Additionally, we also applied the previously mentioned robust methods to these data to obtain a comprehensive analysis of our algorithm's performance. In a scenario featuring 900 samples, each comprising 7 characteristics. For training and testing purposes, 75% of these samples were randomly selected for training, with the remaining used for testing. We consider different rate of corrupted and missing values aiming to evaluate the proposed algorithm ability to handle different levels of outliers & incompleteness. Our experiments were designed to evaluate the performance of different algorithms under various conditions of missing data and outliers.

The results from Fig. 3 and 4 demonstrate not only the effectiveness of *REM-LR* in handling missing data and outliers but also its ability to outperform traditional machine learning algorithms such as *RF*, *KNN*, *SVM*, and even traditional robust algorithms for logistic regression. Specifically, the Area Under the Curve (AUC) is a widely used metric in the context of binary classification problems, indicating the capacity of distinguishing between positive and negative classes. A higher AUC indicates better performance of the classifier. Meanwhile, the Brier score assesses the mean squared deviations between predicted probabilities and actual outcomes (values of 0 or 1). A lower Brier score indicates greater predictive accuracy. This significant innovation in the SAEM algorithm, embodied by *REM-LR*, opens new perspectives in the field of missing data estimation and robustness in logistic regression, highlighting its potential usefulness in the field of machine learning. These results are further validated in the following subsection, where **Table 5** summarize different performance metrics under various scenarios. Our algorithm consistently maintains superior performance, even as the rate of missing data or outliers increases. This robustness highlights *REM-LR's* reliability in challenging conditions, making it an optimal solution for real-world applications with incomplete or corrupted data.

Table 4: Bias and RMSE for 30% MAR and 30% MCAR with 10% of outliers

	REM-LR	SAEM	R-MICE	R-mean	P-M	BY-M	RSI	R-Full
30% MAR + 10% Outliers								
β_1	0.0118	-0.1575	0.0690	0.1010	-0.1996	-0.0977	0.0566	0.0109
β_2	-0.0590	0.2234	-0.0383	0.0222	0.2510	0.3040	0.0515	-0.0151
β_3	0.0132	-0.0050	0.0262	0.0194	-0.0194	0.0324	0.0285	-0.0088
β_4	0.0293	-0.0751	-0.0323	-0.1976	-0.1092	-0.3261	-0.1868	0.0117
β_5	0.0019	0.0111	0.0103	-0.0208	-0.0166	-0.0225	-0.0156	0.0023
β_6	-0.0155	0.0440	0.0130	0.1285	0.0856	0.2086	0.1168	-0.0064
RMSE	**0.0935**	0.1319	0.1145	0.1385	0.1576	0.2127	0.1286	0.0675
30% MCAR + 10% Outliers								
β_1	0.0027	-0.1514	0.0469	0.2038	-0.2204	-0.1310	0.0922	-0.0069
β_2	-0.1105	0.2355	-0.0901	0.0798	0.3437	0.3666	0.1152	0.0027
β_3	0.0339	-0.0073	0.0885	0.0308	0.0054	0.0513	0.0562	-0.0010
β_4	0.0664	-0.0906	-0.0417	-0.4106	-0.3469	-0.4963	-0.3903	0.0069
β_5	0.0077	-0.0004	0.0446	-0.0653	-0.0706	-0.0605	-0.0385	-0.0015
β_6	-0.0410	0.0559	-0.0022	0.2767	0.2557	0.3322	0.2386	-0.0034
RMSE	**0.1264**	0.1378	0.1378	0.2414	0.2522	0.2971	0.2153	0.0680

Table 5: Comparative analysis - mean (standard deviation) - for different MCAR and outlier percentages.

Scenarios	Estimator	AUC	Precision	Accuracy	F1	Brier Score	Specificity
10% MCAR, 10% Outliers	**REM-LR**	**0.93** (0.02)	**0.84** (0.04)	**0.86** (0.02)	**0.87** (0.02)	**0.10** (0.01)	**0.84** (0.04)
	SAEM	0.87 (0.02)	0.76 (0.04)	0.79 (0.03)	0.80 (0.03)	0.15 (0.01)	0.74 (0.04)
	MissForest	0.88 (0.02)	0.77 (0.05)	0.81 (0.03)	0.82 (0.03)	0.14 (0.01)	0.75 (0.05)
	KNN	0.85 (0.02)	0.82 (0.02)	0.85 (0.02)	0.85 (0.02)	0.15 (0.02)	0.81 (0.03)
	RF	0.86 (0.03)	0.83 (0.03)	0.86 (0.02)	0.86 (0.02)	0.14 (0.02)	0.82 (0.04)
	SVM	0.80 (0.02)	0.74 (0.04)	0.79 (0.03)	0.81 (0.02)	0.21 (0.03)	0.69 (0.05)
	P-M	0.88 (0.02)	0.77 (0.04)	0.80 (0.03)	0.81 (0.03)	0.14 (0.01)	0.74 (0.05)
	RSI	0.92 (0.02)	0.83 (0.04)	0.85 (0.02)	0.85 (0.02)	0.11 (0.01)	0.82 (0.04)
30% MCAR, 10% Outliers	**REM-LR**	**0.93** (0.02)	**0.85** (0.03)	**0.87** (0.03)	**0.87** (0.03)	**0.10** (0.02)	**0.85** (0.03)
	SAEM	0.82 (0.03)	0.72 (0.05)	0.74 (0.04)	0.75 (0.04)	0.18 (0.01)	0.71 (0.05)
	MissForest	0.88 (0.02)	0.78 (0.04)	0.80 (0.03)	0.81 (0.03)	0.14 (0.01)	0.76 (0.05)
	KNN	0.84 (0.02)	0.82 (0.03)	0.84 (0.02)	0.85 (0.02)	0.16 (0.02)	0.81 (0.04)
	RF	0.85 (0.03)	0.83 (0.04)	0.85 (0.03)	0.85 (0.03)	0.15 (0.03)	0.82 (0.05)
	SVM	0.79 (0.03)	0.74 (0.04)	0.79 (0.03)	0.81 (0.03)	0.21 (0.03)	0.69 (0.06)
	P-M	0.89 (0.03)	0.78 (0.05)	0.81 (0.04)	0.82 (0.03)	0.13 (0.02)	0.76 (0.06)
	RSI	0.92 (0.02)	0.83 (0.05)	0.85 (0.03)	0.86 (0.03)	0.11 (0.02)	0.81 (0.05)
10% MCAR, 40% Outliers	**REM-LR**	**0.91** (0.03)	**0.83** (0.04)	**0.84** (0.04)	**0.84** (0.05)	**0.12** (0.03)	**0.83** (0.04)
	SAEM	0.82 (0.02)	0.72 (0.04)	0.74 (0.03)	0.75 (0.04)	0.20 (0.01)	0.69 (0.08)
	MissForest	0.87 (0.02)	0.76 (0.03)	0.79 (0.02)	0.80 (0.02)	0.18 (0.01)	0.73 (0.05)
	KNN	0.84 (0.02)	0.82 (0.03)	0.84 (0.02)	0.85 (0.02)	0.16 (0.02)	0.81 (0.04)
	RF	0.84 (0.02)	0.83 (0.04)	0.84 (0.02)	0.85 (0.01)	0.16 (0.02)	0.82 (0.04)
	SVM	0.75 (0.03)	0.70 (0.05)	0.75 (0.03)	0.79 (0.03)	0.25 (0.03)	0.59 (0.11)
	P-M	0.86 (0.02)	0.74 (0.05)	0.77 (0.03)	0.79 (0.03)	0.17 (0.01)	0.69 (0.07)
	RSI	0.47 (0.29)	0.50 (0.08)	0.50 (0.03)	0.67 (0.07)	0.45 (0.08)	0.45 (0.51)
50% MCAR, 10% Outliers	**REM-LR**	**0.93** (0.02)	**0.85** (0.04)	**0.86** (0.03)	**0.86** (0.03)	**0.11** (0.01)	**0.84** (0.04)
	SAEM	0.77 (0.03)	0.68 (0.04)	0.69 (0.04)	0.70 (0.04)	0.20 (0.01)	0.65 (0.06)
	MissForest	0.89 (0.02)	0.78 (0.04)	0.81 (0.03)	0.82 (0.02)	0.14 (0.01)	0.75 (0.06)
	KNN	0.83 (0.03)	0.80 (0.04)	0.83 (0.03)	0.84 (0.02)	0.17 (0.03)	0.78 (0.05)
	RF	0.84 (0.02)	0.82 (0.03)	0.84 (0.02)	0.85 (0.02)	0.16 (0.02)	0.81 (0.04)
	SVM	0.80 (0.04)	0.74 (0.04)	0.80 (0.03)	0.82 (0.03)	0.20 (0.03)	0.68 (0.07)
	P-M	0.88 (0.02)	0.77 (0.04)	0.80 (0.03)	0.81 (0.03)	0.14 (0.01)	0.76 (0.05)
	RSI	0.90 (0.01)	0.82 (0.03)	0.84 (0.02)	0.84 (0.02)	0.12 (0.01)	0.80 (0.04)

Table 6: Performance indicators

Indicator	**Meaning and Formula**
AUC (Area Under the Curve)	Measures the model's ability to distinguish between positive and negative classes.
Precision	Proportion of positive predictions that are actually correct: Precision = $\frac{TP}{TP+FP}$
Accuracy	Proportion of correct predictions (both positive and negative) out of all predictions: Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
F1 Score	Harmonic mean of precision and recall: F1 = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$
Brier Score	Mean squared difference between predicted probabilities and actual outcomes: Brier Score = $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{p}_i)^2$
Specificity	Proportion of true negative predictions out of all actual negatives: Specificity = $\frac{TN}{TN+FP}$
Recall	Proportion of actual positives correctly identified by the model: Recall = $\frac{TP}{TP+FN}$

5. Conclusion

In this article, we addressed the challenge of inference and binary classification with partially incomplete and corrupted covariates. Through the development of a new inference approach integrating expectation-maximization-like algorithm, we presented a novel algroithm design for dealing with missing data in the presence of outliers in complex data structures. Simulations demonstrated the effectiveness and robustness of the proposed method under various scenarios of missing data and outlier model specifications. Future avenues of research could explore the extensions of our methodology and further refinements to increase its utility in various statistical modeling contexts, particularly for dealing with missing data in the missing not at random (MNAR) scenario.

References

[1] Jason W. Osborne. *Best Practices in Logistic Regression*. Sage Publications, 2014.

[2] R. Zhou, J. Liu, S. Kumar and D. Palomar. Student's t var modeling with missing data via stochastic em and gibbs sampling. *IEEE Transactions on Signal Processing*, 68:6198–6211, 2020.

[3] J. Liu, S. Kumar and D. P. Palomar. Parameter estimation of heavy-tailed ar model with missing data via stochastic em. *IEEE Transactions on Signal Processing*, 67(8):2159–2172, 2019.

[4] A.M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma. Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts. *Signal Processing Magazine, IEEE*, 29(4):61 –80, july 2012.

[5] J. Liu and D. P. Palomar. Regularized robust estimation of mean and covariance matrix for incomplete data. *Signal Processing*, pages 278–291, 1995.

[6] Wei Jiang, Julie Josse, Marc Lavielle, TraumaBase Group, et al. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

[7] Nguyen Viet Dung, Nguyen Linh Trung, Karim Abed-Meraim, et al. Robust subspace tracking with missing data and outliers: Novel algorithm with convergence guarantee. *IEEE Transactions on Signal Processing*, 69:2070–2085, 2021.

[8] A Hippert Ferrer, Mohammed Nabil El Korso, Arnaud Breloy, and Guillaume Ginolhac. Robust mean and covariance matrix estimation under heterogeneous mixed-effects model with missing values. *Signal Processing*, 188:108195, 2021.

[9] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing*, 60(11):5597–5625, Nov 2012.

[10] J-P. Delmas, M. N. El Korso, F. Pascal, and S. Fortunati. *Elliptically Symmetric Distributions in Signal Processing and Machine Learning*, volume 4. Springer Nature, New York, 2024.

[11] Raymond J Carroll and Shane Pederson. On robustness in the logistic regression model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(3):693–706, 1993.

[12] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. *Advances in neural information processing systems*, 27, 2014.

[13] Ana M Bianco and Víctor J Yohai. *Robust estimation in the logistic regression model*. Springer, 1996.

[14] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[15] Chuanhai Liu and Donald B Rubin. Ml estimation of the t distribution using em and its extensions, ecm and ecme. *Statistica Sinica*, pages 19–39, 1995.

[16] X. Zhang, M. N. El Korso, and M. Pesavento. MIMO radar target localization and performance evaluation under SIRP clutter. *Signal Processing Journal, Elsevier*, 130(1):217–232, 2017.

[17] Roderick J. A. Little Kenneth L. Lange and Jeremy M. G. Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

[18] Lawrence D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, 1986.

[19] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128, 1999.

[20] Ricardo Maronna, Oscar Bustos, and Victor Yohai. Bias-and efficiency-robustness of general m-estimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation: Proceedings of a Workshop held in Heidelberg, April 2–4, 1979*, pages 91–116. Springer, 2006.

[21] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[22] B. Mériaux, C. Ren, A. Breloy, M. N. El Korso and P. Forster. Matched and mismatched estimation of kronecker product of linearly structured scatter matrices under elliptical distributions. *IEEE Transactions on Signal Processing*, 69(1):603–616, 2021.

[23] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[24] Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[25] Daryl Pregibon. Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, pages 485–498, 1982.

[26] İlkay Çınar, Murat Koklu, and Şakir Taşdemir. Classification of raisin grains using machine vision and artificial intelligence methods. *Gazi Mühendislik Bilimleri Dergisi*, 6(3):200–209, 2020.

[27] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

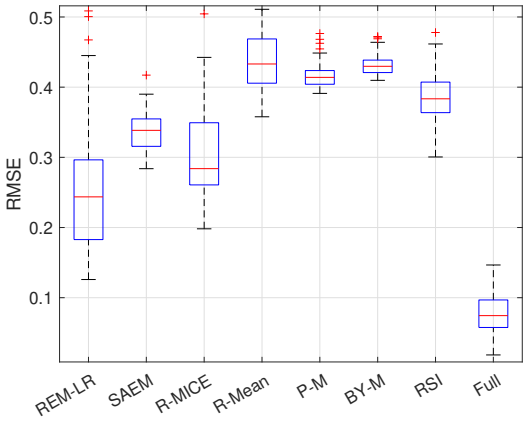

Figure 1: Comparison of the overall RMSE of β with 30% missing data (self-masking MNAR) and 10% outliers for different estimators.

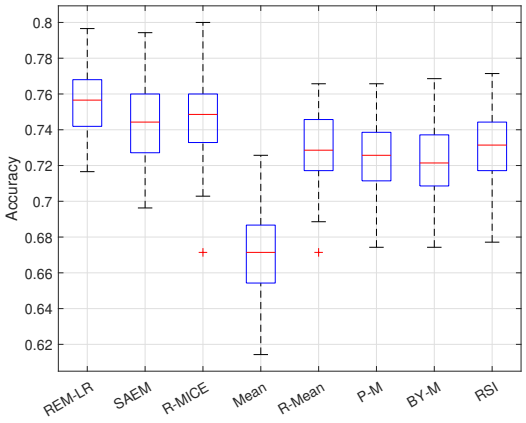

Figure 2: Comparative analysis of accuracy distributions on test sets for 30% MCAR and 10% outliers.

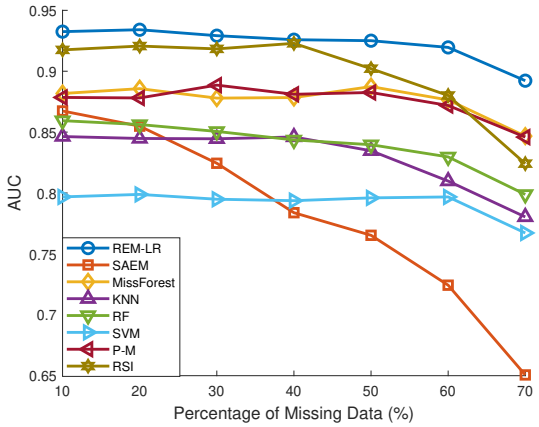

Figure 3: Mean Area Under the Curve (AUC) for 20 Monte Carlo Runs under Different Rates of Missing Data with a step of 10% and 10% of outliers for Each Estimator.

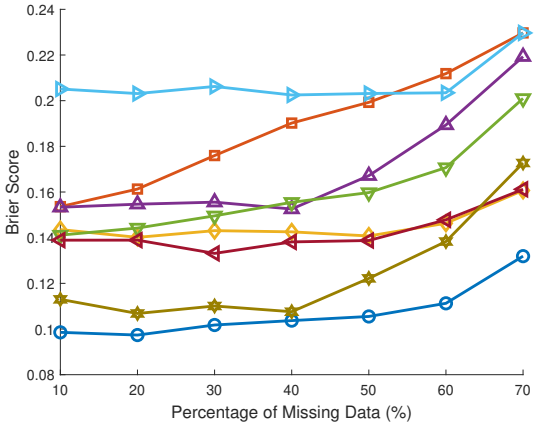

Figure 4: Mean Brier score for 20 Monte Carlo Runs under Different Rates of Missing Data with a step of 10% and 10% of outliers for Each Estimator.