Stochastic Approximation EM for Logistic Regression with Missing Values

Wei Jiang, Julie Josse, Marc Lavielle
CMAP, École Polytechnique  Inria XPOP
1 Background

public health  polytraumatized patients  major trauma

modeling with missing data

misaem
**Traumabase** data:
7495 trauma patients + 244 measurements (quantitative & categorical).

1. Develop the models to help the emergency doctors to take decisions.

<table>
<thead>
<tr>
<th>Type of Accident</th>
<th>Age</th>
<th>Sex</th>
<th>Blood pressure</th>
<th>Lactate</th>
<th>Hemorrhagic shock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falling</td>
<td>50</td>
<td>M</td>
<td>140</td>
<td>NM</td>
<td>Yes</td>
</tr>
<tr>
<td>Fire</td>
<td>28</td>
<td>F</td>
<td>NR</td>
<td>4.8</td>
<td>No</td>
</tr>
<tr>
<td>Knife</td>
<td>30</td>
<td>M</td>
<td>120</td>
<td>1.2</td>
<td>No</td>
</tr>
<tr>
<td>Traffic accident</td>
<td>23</td>
<td>M</td>
<td>110</td>
<td>3.6</td>
<td>No</td>
</tr>
<tr>
<td>Knife</td>
<td>33</td>
<td>M</td>
<td>106</td>
<td>NM</td>
<td>No</td>
</tr>
<tr>
<td>Traffic accident</td>
<td>58</td>
<td>F</td>
<td>150</td>
<td>NM</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Measurements of patients $\xrightarrow{\text{Predict}}$ Hemorrhagic shock

$X_{\text{mixed}} \xrightarrow{\text{Logistic regression}} Y = 0, 1$

**Challenge:** Modeling with **missing data**

(NA = Not Applicable, NM = Not Made, NR = Not Recorded)
State of the art to handle missing values

- Complete case \(\Rightarrow\) loss of information \(\times\)
- Imputation (mice, missMDA, missForest)
- Algorithm Expectation-Maximization to get the maximum likelihood estimators + other algorithms to get the variances
  \(\Rightarrow\) Natural model selection procedure!
  \(\Rightarrow\) Difficult to establish?
  \(\Rightarrow\) Not many implementations, even for simple models.
2 Logistic regression with missing data

SAEM algorithm to estimate the parameters
Estimate variance
Model selection procedure

misaem
Logistic regression model

\[ x = (x_{ij}) \text{ a } n \times p \text{ matrix of quantitative covariates} \]
\[ y = (y_i) \text{ an } n\text{-vector of binary responses } \{0, 1\} \]

**Logistic regression model**

\[
P(y_i = 1|x_i; \beta) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}
\]

**Covariates**

\[ x_i \sim \text{i.i.d. } \mathcal{N}_p(\mu, \Sigma) \]

**Log-likelihood** for complete-data with the set of parameters \( \theta = (\mu, \Sigma, \beta) \)

\[
\mathcal{L}(\theta; x, y) = \sum_{i=1}^{n} \left( \log(p(y_i|x_i; \beta)) + \log(p(x_i; \mu, \Sigma)) \right).
\]
**EM algorithm with missing data**

**Assumption:** Missing data are **Missing at Random**.

**Decomposition:** \( x = (x_{\text{obs}}, x_{\text{mis}}) \).

**Aim:** \( \arg \max \mathcal{L}(\theta; x_{\text{obs}}, y) = \int \mathcal{L}(\theta; x, y) dx_{\text{mis}} \).

**EM:**

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

  \[
  Q_k(\theta) = \mathbb{E}[\mathcal{L}(\theta; x, y) | x_{\text{obs}}, y; \theta_{k-1}] = \int \mathcal{L}(\theta; x, y) p(x_{\text{mis}} | x_{\text{obs}}, y; \theta_{k-1}) dx_{\text{mis}}.
  \]

- **M-step:** Update the estimation of \( \theta \):

  \[
  \theta_k = \arg \max_{\theta} Q_k(\theta).
  \]

**Unfeasible computation of expectation!**
(book, Lavielle 2014)

- **Simulation:** For \( i = 1, 2, \ldots, n \), draw one sample \( x_{i,\text{mis}}^{(k)} \) from target distribution

\[
p(x_{i,\text{mis}}|x_{i,\text{obs}}, y_i; \theta_{k-1}).
\]

- **Stochastic approximation:** Update the function \( Q \)

\[
Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left( \mathcal{L}(\theta; x_{\text{obs}}, x_{\text{mis}}^{(k)}, y) - Q_{k-1}(\theta) \right),
\]

where \( (\gamma_k) \) is a decreasing sequence of positive numbers.

- **Maximization:** \( \theta_k = \arg \max_{\theta} Q_k(\theta) \).
Metropolis-Hastings algorithm

Target distribution

\[ f_i(x_{i,mis}) = p(x_{i,mis}|x_{i,obs}, y_i; \theta) \]
\[ \propto p(y_i|x_i; \beta) p(x_{i,mis}|x_{i,obs}; \mu, \Sigma). \]

Proposal distribution

\[ g_i(x_{i,mis}) = p(x_{i,mis}|x_{i,obs}; \mu, \Sigma) \sim \mathcal{N}_p(\mu_i, \Sigma_i) \]
\[ \mu_i = \mu_{mis} + \Sigma_{mis,obs} \Sigma_{obs,obs}^{-1}(x_{i,obs} - \mu_{obs}), \]
\[ \Sigma_i = \Sigma_{mis,mis} - \Sigma_{mis,obs} \Sigma_{obs,obs}^{-1} \Sigma_{obs,mis}. \]

Metropolis:

- \[ z_{i,m}^{(k)} \sim g_i(x_{i,mis}), \; u \sim \mathcal{U}[0, 1] \]
- \[ r = \frac{f_i(z_{i,m}^{(k)})/g_i(z_{i,m}^{(k)})}{f_i(z_{i,m-1}^{(k)})/g_i(z_{i,m-1}^{(k)})} \]
- If \( u < r \), accept \( z_{i,m}^{(k)} \)}
Variance estimation with Louis formula: $\mathcal{I}_{\text{obs}} = \mathcal{I}_{\text{comp}} - \mathcal{I}_{\text{mis}}$

Model selection criterion:

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

Observed likelihood:

\[
p(y_i, x_{i,\text{obs}}; \theta) = \int p(y_i, x_{i,\text{obs}}|x_{i,\text{mis}}; \theta)p(x_{i,\text{mis}}; \theta)dx_{i,\text{mis}}
\]

\[
= \int p(y_i, x_{i,\text{obs}}|x_{i,\text{mis}}; \theta)\frac{p(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})}g_i(x_{i,\text{mis}})dx_{i,\text{mis}}
\]

\[
= \mathbb{E}_{g_i} \left( p(y_i, x_{i,\text{obs}}|x_{i,\text{mis}}; \theta)\frac{p(x_{i,\text{mis}}; \theta)}{g_i(x_{i,\text{mis}})} \right).
\]

Sample from $g_i$ (proposal distribution) $\Rightarrow$ Empirical mean.
Package: misaem

library(devtools); install_github("wjiang94/misaem")
library(misaem)
list.saem = miss.saem(X.obs, y, maxruns = 500, tol_em = 1e-07)

Arguments:

- X.obs: Design matrix with missingness $n \times p$;
- y: Response vector $n \times 1$.

Return a list with components:

- beta: Estimated $\beta$;
- var: Estimated variance for estimated parameters;
- logl: Observed log-likelihood.
Simulation study

SAEM behavior
model selection effects

Comparison with other methods

bias of estimation
Comparison with competitors: bias

\[ x: p = 5, n = 1000 / n = 10000 \Rightarrow y \in \{0, 1\}, \]
10% missingness. Repeat 1000 times for each setting.

<table>
<thead>
<tr>
<th></th>
<th>(n = 1000)</th>
<th>(n = 10000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mice</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAEM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**no NA**: classical estimation on original dataset without NA;
**CC**: complete case analysis method;
**mice**: multiple imputation implemented by package *mice*.
Comparison with competitors: confidence interval

Table: Coverage (%) for \( n = 10000 \), calculated over 1000 simulations.

<table>
<thead>
<tr>
<th>parameter</th>
<th>no NA</th>
<th>CC</th>
<th>mice</th>
<th>SAEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>95.2</td>
<td>94.4</td>
<td>95.2</td>
<td>94.9</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>96.0</td>
<td>94.7</td>
<td>93.9</td>
<td>95.1</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>95.5</td>
<td>94.6</td>
<td>94.0</td>
<td>94.3</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>94.9</td>
<td>94.3</td>
<td>86.5</td>
<td>94.7</td>
</tr>
<tr>
<td>( \beta_4 )</td>
<td>94.6</td>
<td>94.2</td>
<td>96.2</td>
<td>95.4</td>
</tr>
<tr>
<td>( \beta_5 )</td>
<td>95.9</td>
<td>94.4</td>
<td>89.6</td>
<td>94.7</td>
</tr>
</tbody>
</table>

Figure: Length of confidence interval.
3 Comparison with competitors: execution time

Table: Comparison of execution time (in seconds) with \( n = 1000 \) calculated over 1000 simulations.

<table>
<thead>
<tr>
<th></th>
<th>no NA</th>
<th>MCEM</th>
<th>mice</th>
<th>SAEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>2.87 \times 10^{-3}</td>
<td>492</td>
<td>0.64</td>
<td>9.96</td>
</tr>
<tr>
<td>mean</td>
<td>4.65 \times 10^{-3}</td>
<td>773</td>
<td>0.70</td>
<td>13.50</td>
</tr>
<tr>
<td>max</td>
<td>43.50 \times 10^{-3}</td>
<td>1077</td>
<td>0.76</td>
<td>16.79</td>
</tr>
</tbody>
</table>

Extract of code (MCMC):

```r
for (m in 1:nmcmc){
  xina.c <- mi + rnorm(njna)%*%chol(0i)
  if (y[i]==1)
    alpha <- (1+exp(-sum(xina*betana))/cobs)/(1+exp(-sum(xina.c*betana))/cobs)
  else
    alpha <- (1+exp(sum(xina*betana))*cobs)/(1+exp(sum(xina.c*betana))*cobs)
  if (runif(1) < alpha){xina <- xina.c}
}
```
# Model selection results

Table: Percentage of times that different criterion selects the correct true model (C), overfit (O), and underfit (U).

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Non-Correlated</th>
<th>Correlated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>O</td>
</tr>
<tr>
<td>$BIC_{obs}$</td>
<td>92</td>
<td>3</td>
</tr>
<tr>
<td>$BIC_{orig}$</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>$BIC_{cc}$</td>
<td>79</td>
<td>1</td>
</tr>
</tbody>
</table>

Extract of codes:

```r
for (j in 1:(nrow(subsets)-1)){
  variables = subsets[j,]
  pos_var=which(variables==1)
  nb.x = sum(variables)
  nb.para = (nb.x + 1) + p + p*p
  list.saem.subset=miss.saem(X.obs,y,pos_var,ll_obs_cal=TRUE)
  BIC[nb,j] = -2*list.saem.subset$ll+ nb.para * log(n)
}
```
4 Application on Traumabase

Exploration of dataset
Cross validation
Risk of severe hemorrhage for Traumabase
Predictive performance
Exploration of dataset

Data preprocessing ⇒ **6384 patients**.
Clinical experience ⇒ **14 influential quantitative measurements**
Missingness: 0 - **60%**.

**Figure:** Percentage of missingness in each variable.
Exploration of dataset

Two observations resulting in a very small value of observed log-likelihood:
For the 3302nd patient, the calculation of BMI is wrong.
For 1144th patient, the values of Weight (200 kg) and Height (100 cm) have a large possibility to be wrong.
4 Estimation & interpretation

Random split: training set (80%) + test set (20%)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Estimate (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.52 (0.59)</td>
</tr>
<tr>
<td>Age</td>
<td>0.011 (0.0033)</td>
</tr>
<tr>
<td>Glasgow.moteur</td>
<td>-0.16 (0.036)</td>
</tr>
<tr>
<td>FC.max</td>
<td>0.026 (0.0025)</td>
</tr>
<tr>
<td>Hemocue.init</td>
<td>-0.23 (0.031)</td>
</tr>
<tr>
<td>RT.cristalloides</td>
<td>0.00090 (0.00010)</td>
</tr>
<tr>
<td>RT.colloides</td>
<td>0.0019 (0.00021)</td>
</tr>
<tr>
<td>SD.min</td>
<td>-0.025 (0.0050)</td>
</tr>
<tr>
<td>SD.SMUR</td>
<td>-0.021 (0.0056)</td>
</tr>
</tbody>
</table>

- The more one bleed, the lower the HemoCu is, and the more the blood will be transfused. Then the more likely one will end up in a hemorrhagic shock.
4 Predictive performance: comparison

Random split: training set (80%) + test set (20%) (repeated 15 times)

Table: Comparison of the median of the predictive performances (values are multiplied by 100) of different methods dealing with missing data.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SAEM</th>
<th>impPCA</th>
<th>impMean</th>
<th>mice</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>86.70</td>
<td>86.67</td>
<td>86.62</td>
<td>86.62</td>
</tr>
<tr>
<td>Accuracy</td>
<td>83.23</td>
<td>81.96</td>
<td>82.74</td>
<td>83.54</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>78.26</td>
<td>77.59</td>
<td>76.86</td>
<td>76.29</td>
</tr>
<tr>
<td>Specificity</td>
<td>83.70</td>
<td>82.21</td>
<td>83.15</td>
<td>84.58</td>
</tr>
<tr>
<td>Precision</td>
<td>30.56</td>
<td>30.68</td>
<td>30.97</td>
<td>32.88</td>
</tr>
</tbody>
</table>
5 Conclusion

- SAEM for logistic regression with missingness leads to unbiased estimation and a more reasonable coverage of confidence interval;
- Model selection by criterion BIC with missing data can be well performed;
- R package misaem: github.com/wjiang94/miSAEM_logReg
- arXiv:1805.04602

1. Deal with both quantitative and categorical data;
2. Deal with both MAR and MNAR missing values.
3. Causal inference and propensity score analysis.