An implementation of the SAEM algorithm for left-censored data

Raphaël Coudret and Jan Serroyen
(Open Analytics and Janssen Pharmaceutica)

The useR! Conference,
June 30 – July 3, 2015, Aalborg, Denmark
Contents

The SAEM algorithm and our implementation

Models with left-censored observations

Comparison between SAEM and EM

Future work
**Goal:** Find the maximum likelihood estimator for some unknown vector of parameters $\theta$.

**Problem:** The likelihood function $(\theta^* \mapsto f_Y|_{\theta=\theta^*}(y))$ can be difficult to write.

**Problem:** The expectation of the EM algorithm can be difficult to write.
The SAEM algorithm and our implementation

Let $Z$ be some unobserved random vector.

The SAEM algorithm:

- generates $z$ from the distribution of $(Z \mid Y = y, \theta = \hat{\theta}_m)$,  
  $\rightarrow$ S step
- finds $f_{Y,Z \mid \theta = \theta^*}(y,z)$,
- produces a function to optimize leading to $\hat{\theta}_{m+1}$.

Convergence of a stochastic approximation version of the EM algorithm.  
The SAEM algorithm and our implementation

An important requirement for the SAEM algorithm to have pleasing properties is, for $Y, Z|\theta = \theta^*$, to be in the curved exponential family.

This means that:

$$f_{Y,Z|\theta=\theta^*}(y, z) = e^{-\Lambda(\theta^*) + \langle S(y,z), \Phi(\theta^*) \rangle},$$

where $S$ is the minimal sufficient statistic of $(Y', Z')'$. 
The SAEM algorithm and our implementation

In the saemCensoring package, there is an implementation of the SAEM algorithm:

- handling models with left-censored observations,
- that can be compared with the EM algorithm, for a particular model,
- capable of ending after each iteration.
Models with left-censored observations

We consider the following model:

- $\mu \in \mathbb{R}^p$,  
- $\Omega$ is a $p \times p$ diagonal positive-definite matrix,  
- $\phi_i \sim \mathcal{N}(\mu, \Omega)$,  
- $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$,  
- $y_{i,j}^{\text{cens}} = h(\phi_i, t_{i,j}) + \varepsilon_{i,j}$,  
- $y_{i,j}^{\text{obs}} = y_{i,j}^{\text{cens}} \mathbb{1}_{\{y_{i,j}^{\text{cens}} \geq \text{LOQ}\}} + \text{LOQ} \mathbb{1}_{\{y_{i,j}^{\text{cens}} < \text{LOQ}\}}$,  
- all $\phi_i$'s and $\varepsilon_{i,j}$'s are independent.
Models with left-censored observations

We then choose:

$$\theta = (\mu', \omega_1^2, \ldots, \omega_p^2, \sigma^2)'$$

where

$$\Omega = \begin{pmatrix}
\omega_1^2 & 0 & \cdots & 0 \\
0 & \omega_2^2 & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \cdots & 0 & \omega_p^2
\end{pmatrix}.$$

We also choose:

$$Y = (y_{1,1}^{obs}, \ldots, y_{N,n_N}^{obs})'.$$

and

$$Z = (\phi_1', \ldots, \phi_N', y_{1,1}^{cens}, \ldots, y_{N,n_N}^{cens})'.$
Models with left-censored observations

We have that $\mathbf{Y}, \mathbf{Z}|\theta = \theta^*$ is in the curved exponential family.

No assumption about the function $h$.

**BUT**

We did not verify all the assumptions in Delyon et al. (1999).

---

**Coudret, R. (2014).**

**Samson, A., Lavielle, M. and Mentré F. (2006).**
Comparison between SAEM and EM

In the model with left-censored observations, if we choose $p = 1$, $LOQ = -\infty$, and:

$$h(\phi_i, t_{i,j}) = \mathbb{I}\{\phi_i > 1\} - \mathbb{I}\{\phi_i \leq -1\},$$

we can write the equations of:

$$f_{Y|\theta=\theta^*}(y) \quad \text{and} \quad f_{\phi_1,\ldots,\phi_N|Y=y,\theta=\theta^*}(u'_1, \ldots, u'_N),$$

and observe the behaviour of both the SAEM and the EM algorithm.
For the S step, Samson et al. (2006) proposed to generate an observation from the distribution of:

$$\psi = \left((\phi_1', \ldots, \phi_N') \mid Y = y, \theta = \theta^*\right),$$

using a Metropolis-Hastings algorithm.

Since we know the density of this random vector, we can compare it with the estimated density computed from the points simulated using the Metropolis-Hastings algorithm.
Comparison between SAEM and EM

Figure: density of $\psi$ (black) and estimate of this density using 900 points created with the function `generateMissingData` (red).
Comparison between SAEM and EM

While running, the `saem` function can show successive estimates of a parameter in $\theta$.

If the `RGtk2` package is correctly installed, a button allows the user to stop neatly the SAEM algorithm after the current iteration. → Quick results when the estimates do not change.

Possible future features:

- being able to set values for $\hat{\theta}_m$ when the function is running, → explore regions of the parameter space.
- show several figures, for all parameters in $\hat{\theta}_m$. 
We chose $N = 10$, $n_i = 10$ for all $i \in \mathbb{N}^*$ and simulated data using $	heta = (3, 4, 0.25)'$.

We launched 100 times the SAEM and the EM algorithms with this data-set, and we found the following values of $\log \left( f_{Y|\theta = \hat{\theta}_m}(y) \right)$:

- EM algorithm: $-76.4023 \pm 10^{-4}$,
- SAEM algorithm: $-76.4006 \pm 10^{-4}$. 
Future work

Interesting tasks that remain to be completed:

- verify all the assumptions in Delyon et al. (1999),
  - study the consequences for $h$,
  - determine whether $S$ has to be the minimal sufficient statistic,

- compare the saemCensoring package with other implementations,

- find what happens when $\Omega$ is not diagonal,

- improve the graphical user interface.