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

Goal: Find the maximum likelihood estimator for some unknown vector of parameters θ.

Problem: The likelihood function ($\theta^\star \mapsto f_{\mathbf{Y}|\theta=\theta^\star}(y)$) can be difficult to write.

Problem: The expectation of the EM algorithm can be difficult to write.

The SAEM algorithm and our implementation

Let $\mathbf{Z}$ be some unobserved random vector.

The SAEM algorithm:

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

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.

The SAEM algorithm and our implementation

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

This means that:

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

where S is the minimal sufficient statistic of $(\mathbf{Y}', \mathbf{Z}')'$.

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:

- $\mu \in \mathbb{R}^p$,
- Ω is a $p \times p$ diagonal positive-definite matrix,
- $\phi_i \sim \mathcal{N}(\mu, \Omega)$,
- $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$,
- $y_{i,j}^{cens} = h(\phi_i, t_{i,j}) + \varepsilon_{i,j}$,
- $y_{i,j}^{obs} = y_{i,j}^{cens}\ \mathbb{I}_{\{y_{i,j}^{cens} \geq LOQ\}} + LOQ\ \mathbb{I}_{\{y_{i,j}^{cens} < LOQ\}}$,
- all ϕ_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:

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

and

$$\mathbf{Z} = (\phi_1', \ldots, \phi_N', y_{1,1}^{cens}, \ldots, y_{N,n_N}^{cens})'.$$

Models with left-censored observations

We have that $\mathbf{Y}, \mathbf{Z} | \theta = \theta^\star$ 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).
Notes on "Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model".
Technical report, Open Analytics.

Samson, A., Lavielle, M. and Mentré F. (2006).
Extension of the SAEM algorithm to left-censored data in nonlinear mixed-effects model: Application to HIV dynamics model.
Computational Statistics & Data Analysis, 51, 1562–1574.

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:

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

and observe the behaviour of both the SAEM and the EM algorithm.

Comparison between SAEM and EM

For the S step, Samson et al. (2006) proposed to generate an observation from the distribution of:

$$\psi = \Big((\phi_1', \ldots, \phi_N')' | \mathbf{Y} = y, \theta = \theta^\star\Big),$$

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

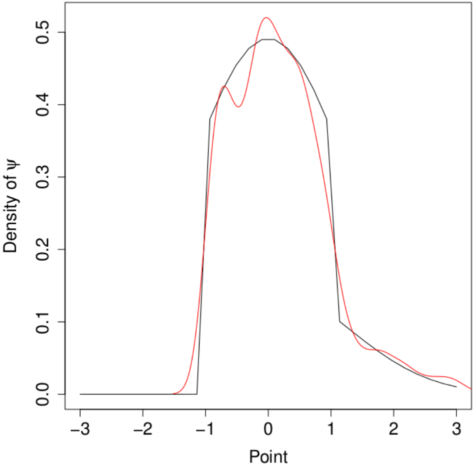

Figure: density of ψ (black) and estimate of this density using 900 points created with the function `generateMissingData` (red).

Comparison between SAEM and EM

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

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

Possible future features:

- being able to set values for $\widehat{\theta}_m$ when the function is running,
  $\rightarrow$ explore regions of the parameter space.
- show several figures, for all parameters in $\widehat{\theta}_m$.

Comparison between SAEM and EM

We chose $N = 10$, $n_i = 10$ for all $i \in \mathbb{N}_N^*$ and simulated data using $\theta = (3, 4, 0.25)'$.

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

- EM algorithm: -76.4023 ± 10^{-4},
- SAEM algorithm: -76.4006 ± 10^{-4}.

Future work

Interesting tasks that remain to be completed:

- verify all the assumptions in Delyon et al. (1999),
  $\rightarrow$ study the consequences for h,
  $\rightarrow$ determine whether S has to be the minimal sufficient statistic,
- compare the `saemCensoring` package with other implementations,
- find what happens when Ω is not diagonal,
- improve the graphical user interface.