# EM algorithm and Simulated EM algorithms

Thierry Kamionka

2020-2021

- The EM algorithm helps to find the Maximum Likelihood Estimator in incomplete data problems.

  Let $Y_1^*, Y_2^*, ..., Y_n^*$ be iid ramdon variables with density $f_\theta(y^*)$. We do not observe $y_i^*$. We observe $y_i = Y_i(Y_i^*)$ (incomplete sampling scheme).
- The EM alogirithm has twos steps : an E-step and an M-step.
- The E-step is the calculation of the conditional expectation of the log-likelihood in complete information given the observed data.
- The M-step is the maximization of the previous quantity.
- These two steps are iterated .

Let $f_\theta(y)$ denote the marginal density of $y$. Let us assume we observe an iid sample $(y_1, \ldots, y_n)$.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_\theta(y_i)$$

where $\theta$ is a vector of parameters.

For instance, $y^* = (y, z)$ and $Y(y^*) = y$ (z can be considered as an unobserved heterogeneity component). For example, $y = y^* \, 1\!\!1[y^* \geq 0]$ (selection).

Let us assume that it is difficult to maximize such a likelihood function. Indeed, the density in the context on an incomplete sample scheme is an expectation. It can be easier to maximize an objective function that involve the expectation of the log of a density.

**EM algorithm** (Dempster, Laird, Rubin, 1977):

Let us consider the following function of the parameters

$$Q(\theta \mid \theta') = \frac{1}{n}\sum_{i=1}^{n} \underset{\theta'}{E}[\log(f_\theta(y_i^*)) \mid y_i].$$

We start we an initial value for the vector of parameters $(\theta' \leftarrow \theta^{(0)})$.

One iteration of the algorithm :

The **E-step**: Calculation of $Q(\theta \mid \theta')$ for $\theta'$ fixed.

The **M-step**: Consists to maximize $Q(\theta \mid \theta')$ with respect to $\theta$.

Then we replace replace $\theta'$ by the new value of $\theta$ obtained from the M-step and we iterate from step k to step k+1 $(\theta' \leftarrow \theta^{(k)})$.

Let $M(\theta') = argmax_\theta \, Q(\theta \mid \theta')$. The MLE of the $\theta$ (namely $\hat{\theta}$) is a **fixed point** of $M(\theta)$ : it is such that $\hat{\theta} = M(\hat{\theta})$.

Let $\theta^{(k)}$ denote the value of $\theta$ obtained at the iteration $k$ of the EM algorithm. The algorithm is such that $L(\theta^{(k)}) \geq L(\theta^{(k-1)})$.

Let $\Omega(\theta, \theta^{(k)}) = L(\theta) - L(\theta^{(k)})$. We are going to show that $\Omega(\theta^{(k+1)}, \theta^{(k)}) \geq 0$.

Let us assume to simplify the presentation of the proof that $y^* = (y, z)$ and the variable $z$ is discrete.

Sketch of proof: Let

$$\Omega(\theta, \theta^{(k)}) = \frac{1}{n} \sum_{i=1}^{n} \log(f_\theta(y_i)) - \frac{1}{n} \sum_{i=1}^{n} \log(f_{\theta^{(k)}}(y_i))$$

where $\log(f_\theta(y)) = \sum_z f_\theta(y \mid z) \pi_\theta(z)$.

$$\Omega(\theta,\theta^{(k)}) = \frac{1}{n}\sum_{i=1}^{n}\log(\sum_{z} f_\theta(y_i \mid z)\pi_\theta(z)) - \frac{1}{n}\sum_{i=1}^{n}\log(f_{\theta^{(k)}}(y_i))$$

$$= \frac{1}{n}\sum_{i=1}^{n}\log(\sum_{z}\frac{f_\theta(y_i|z)\pi_\theta(z)}{Prob[z|y_i;\theta^{(k)}]}Prob[z|y_i;\theta^{(k)}])$$

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(f_{\theta^{(k)}}(y_i))$$

As $-\log(x)$ is a convex function, the **Jensen's inequality** allows to write

$$\Omega(\theta,\theta^{(k)}) \geq \frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(\frac{f_\theta(y_i \mid z)\pi_\theta(z)}{Prob[z|y_i;\theta^{(k)}]})$$

$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(f_{\theta^{(k)}}(y_i))$$

$$\Omega(\theta,\theta^{(k)}) \geq \frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(\frac{f_\theta(y_i \mid z)\pi_\theta(z)}{Prob[z|y_i;\theta^{(k)}]f_{\theta^{(k)}}(y_i)})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(\frac{f_\theta(y_i;z)}{f_{\theta^{(k)}}(y_i;z)}) \equiv \omega(\theta,\theta^{(k)})$$

Let us remark that

$$\begin{aligned}
\theta^{(k+1)} &= \underset{\theta}{argmax}\,\omega(\theta,\theta^{(k)}) \\
&= \underset{\theta}{argmax}\,\frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(\frac{f_\theta(y_i;z)}{f_{\theta^{(k)}}(y_i;z)}) \\
&= \underset{\theta}{argmax}\,\frac{1}{n}\sum_{i=1}^{n}\sum_{z} Prob[z|y_i;\theta^{(k)}]\log(f_\theta(y_i;z))
\end{aligned}$$

$$\begin{aligned}\theta^{(k+1)} &= \underset{\theta}{argmax}\, \frac{1}{n}\sum_{i=1}^{n}\sum_{z}\frac{\pi_{\theta^{(k)}}(z)f_{\theta^{(k)}}(y_i;z)}{\sum_z \pi_{\theta^{(k)}}(z)f_{\theta^{(k)}}(y_i;z)}\log(f_\theta(y_i;z))\\ &= \underset{\theta}{argmax}\, \frac{1}{n}\sum_{i=1}^{n}\underset{\theta^{(k)}}{E}\left[\log(f_\theta(y_i;z)\mid y_i\right]\end{aligned}$$

We have shown that $\log(L(\theta^{(k+1)})) - \log(L(\theta^{(k)})) = \Omega(\theta^{(k+1)},\theta^{(k)}) \geq \omega(\theta^{(k+1)},\theta^{(k)}) \geq \omega(\theta^{(k)},\theta^{(k)}) = 0$. QED

An example : EM algorithm for the exponential distribution and right censoring.

Remark : this distribution belongs to the exponential family (see Gouriéroux, Monfort, Trognon, 1984).
The density (complete observation scheme) :

$$f_\theta(y^*) = \lambda exp(-\lambda y^*)$$

for all $y^* > 0$. The cumulative distribution function is

$$Prob[Y^* \leq a] = 1 - exp(-\lambda a)$$

**Incomplete sampling scheme** : We observe $y = Y(y^*)$ such that

$$y = y^* \text{ if } y^* < a$$

$$y = a \text{ if } y^* \geq a > 0$$

## Example

Interpretation : $Y^*$ is a duration and this duration can be right censored.

The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} (\lambda exp(-\lambda y_i^*))^{\mathit{1}\{y_i<a\}} (\exp(-\lambda\ a))^{\mathit{1}\{y_i=a\}}$$

In this example, the likelihood function can be maximized with respect to the parameter $\lambda$ ($\lambda > 0$).

The distribution of $Y^*$ given $Y = y$ is given by the density

$$\kappa_\theta(y^* \mid y) = 1 \text{ if } y < a \text{ and } y^* = y,$$

$$\kappa_\theta(y^* \mid y) = 0 \text{ if } y < a \text{ and } y^* \neq y,$$

$$\kappa_\theta(y^* \mid y) = \lambda exp(-\lambda y^*)/exp(-\lambda a), \text{ if } y = a,$$

The E-step of the EM algorithm consists in calculating

$$Q(\theta \mid \theta') = \frac{1}{n}\sum_{i=1}^{n} E_{\theta'}[\log(f_\theta(Y_i^*)) \mid y_i]$$

for $\theta'$ fixed. A fixed point of the function $M(\theta)$ is such that

$$\frac{\partial}{\partial\theta} Q(\theta \mid \theta') \mid_{\theta=\theta'} = 0$$

this is equivalent to find $\theta'$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} E_{\theta'}[\log(f_\theta(Y_i^*)) \mid y_i] \mid_{\theta=\theta'} = 0$$

In order to find the root of this previous equation, we operate iteratively. We start with a value $\theta'$ and the next value is such that

$$\frac{\partial}{\partial\theta} Q(\theta \mid \theta') = 0$$

Example (continued):

$$Q(\theta \mid \theta') = \frac{1}{n} \sum_{i:y_i<a} E_{\lambda'}[\log(\lambda exp(-\lambda Y_i^*)) \mid y_i]$$

$$+ \frac{1}{n} \sum_{i:y_i=a} E_{\lambda'}[\log(\lambda e^{-\lambda Y_i^*}) \mid y_i]$$

$$Q(\theta \mid \theta') = \frac{1}{n} \sum_{i:y_i<a} \log(\lambda exp(-\lambda y_i))$$

$$+ \frac{1}{n} \sum_{i:y_i=a} \int_a^\infty \log(\lambda e^{-\lambda y}) \;\; \lambda' e^{-\lambda' y} / e^{-\lambda' a} \, d \, y$$

## Example

$$Q(\theta \mid \theta') = \frac{1}{n} \sum_{i:y_i<a} (\log(\lambda) - \lambda y_i)$$

$$+ \frac{1}{n} \sum_{i:y_i=a} [\log(\lambda) - \lambda \int_a^{\infty} y \ \ \lambda' e^{-\lambda' y} / e^{-\lambda' a} \, d \ y]$$

$$Q(\theta \mid \theta') = \frac{1}{n} \sum_{i:y_i<a} (\log(\lambda) - \lambda y_i)$$

$$+ \frac{1}{n} \sum_{i:y_i=a} [\log(\lambda) - \lambda(a + \frac{1}{\lambda'})]$$

## Example

Let us assume that $\lambda = \lambda' = \hat{\lambda}_n$ then we obtain

$$\frac{1}{n} \sum_{i:y_i<a} [\frac{1}{\hat{\lambda}_n} - y_i] + \frac{1}{n} \sum_{i:y_i=a} [\frac{1}{\hat{\lambda}_n} - (a + \frac{1}{\hat{\lambda}_n})] = 0$$

$$\frac{1}{\hat{\lambda}_n} \sum_{i:y_i<a} 1 = \sum_{i:y_i<a} y_i + \sum_{i:y_i=a} a$$

$\hat{\lambda}_n$ is the fixed point of $M(\theta)$. It is the **Maximum Likelihood Estimator** of $\theta = \lambda$.

The likelihood function is

$$L(\theta) = \prod_{i:y_i<a} \lambda \exp(-\lambda y_i) \prod_{i:y_i=a} \exp(-\lambda a)$$

The log-likelihood function is

$$\log(L(\theta)) = \sum_{i:y_i<a} \log(\lambda) - \sum_{i:y_i<a} \lambda y_i - \sum_{i:y_i=a} \lambda a$$

The first order condition is

$$\frac{1}{\hat{\lambda}_n} \sum_{i:y_i<a} 1 = \sum_{i:y_i<a} y_i + \sum_{i:y_i=a} a$$

Consequently one can verify that the EM algorithm allows to obtain at the **stationarity of the Markov chain** the value of the MLE.

The EM algorithm can be extended to the cases such that the expectation to be evaluated in the Maximization step **do not have any analytical expression.**

## Example (continued): Description of the Markov Chain

Let us consider the first order condition associated to $Q(\lambda \mid \lambda')$,

$$\frac{1}{n}\sum_{i:y_i<a}\frac{1}{\lambda}-\frac{1}{n}\sum_{i:y_i<a}y_i+$$

$$+\frac{1}{n}\sum_{i:y_i=a}\frac{1}{\lambda}-\frac{1}{n}\sum_{i:y_i=a}(a+\frac{1}{\lambda'})=0$$

Consequently,

$$\frac{1}{\lambda}=\frac{1}{n}\sum_{i:y_i<a}y_i+\frac{1}{n}\sum_{i:y_i=a}(a+\frac{1}{\lambda'}).$$

Then let us fix $\lambda' \leftarrow \lambda$ and we iterate.

Let , $y_{ik}^*$, $k = 1, \ldots, H$, be a random draw from the conditional distribution $\kappa_{\theta'}(y^* \mid y)$. An **unbiaised** estimate of $Q(\theta \mid \theta')$ is given by

$$\hat{Q}(\theta \mid \theta') = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{H} \sum_{k=1}^{H} \log(f_\theta(y_{ik}^*))$$

The draws $y_{ik}^*$ are often obtained using independent uniform random numbers $y_{ik}^* = \psi(u_{ik} \mid y_i, \theta')$ (or vectors of independent uniform random numbers). The expression of $\psi$ is related to the one of $\kappa_{\theta'}(y^* \mid y)$.

We can have two different Simulated EM algorithms :

- If we draw new independent variables $u_{ik}$ in each iteration, the sequence of $\theta$ obtained from the algorithm is an ergodic Markov chain. It is the **StEM algorithm** (Nielsen 2000).
- If we use the same independent variables $u_{ik}$ for all iterations, it is the **SimEM algorithm** (Nielsen 2000).

## Example (continued):

$$\kappa_{\theta'}(y^* \mid y) = \frac{\lambda' \exp(-\lambda' \, y^*)}{\exp(-\lambda' a)}$$

where $y^* \geq a$ $(y = a)$.
The conditonal cumulative distribution function is

$$\eta_{\theta'}(y^*) = 1 - \frac{\exp(-\lambda' y^*)}{\exp(-\lambda' a)}$$

Let $u_{ik}$ denote an independent random draw in the uniform distribution. Then $y^*_{ik} = \eta^{-1}_{\lambda'}(u_{ik}) = \psi(u_{ik} \mid a; \lambda')$ is an independent random draw in the conditional distribution $\kappa_{\theta'}(y^* \mid y)$.
One can verify that

$$y^*_{ik} = \eta^{-1}_{\lambda'}(u_{ik}) = \frac{1}{\lambda'}[a\,\lambda' - \log(1 - u_{ik})] = a - \frac{1}{\lambda'}\log(1 - u_{ik})$$

**SimEM algorithm**: The drawings $u_{ik}$, $k = 1, \ldots, H$ **are constant** through the iterations of the algorithm. As the expression of the objective function $\hat{Q}(\theta \mid \theta')$ do not vary for $\theta$ and $\theta'$ fixed, the **E-step** is very simple and consists to replace the former value of $\theta'$ by a new one (the argmax of the previous **maximization step**).

We start with a value of $\theta'$ and we solve the equation (f.o.c. of the maximization step)

$$\frac{\partial}{\partial\theta}\hat{Q}(\theta \mid \theta') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{k=1}^{H}\frac{\partial}{\partial\theta}log(f_\theta(\psi(u_{ik} \mid y_i, \theta')) = 0$$

with respect to $\theta$. We assume that we can permute the order of differentiation and of expectation (i.e. the derivative of the expectation is the expectation of the derivative).

This root $\theta$ is then used to replace the value of $\theta'$ for the next iteration of the algorithm.

## **SimEM algorithm**: Example (continued)

$$\hat{Q}(\theta \mid \theta') = \frac{1}{n} \sum_{i:y_i<a} (\log(\lambda) - \lambda y_i)$$

$$+ \frac{1}{n} \sum_{i:y_i=a} [\log(\lambda) - \lambda \, \frac{1}{H} \sum_{k=1}^{H} \, y^*_{ik} \;\; ]$$

where $y^*_{ik} = \psi(u_{ik} \mid y, \theta') = \eta^{-1}_{\lambda'}(u_{ik}) = \frac{1}{\lambda'}[a\, \lambda' - \log(1 - u_{ik})]$.
We have to find the root of

$$\frac{\partial}{\partial \theta} \hat{Q}(\theta \mid \theta') = \frac{1}{n} \sum_{i:y_i<a} (\frac{1}{\lambda} - y_i)$$

$$+ \frac{1}{n} \sum_{i:y_i=a} [\frac{1}{\lambda} - \frac{1}{H} \sum_{k=1}^{H} \, y^*_{ik} \;\; ] = 0$$

The root $\lambda$ is such that

$$\frac{1}{\lambda} = \frac{1}{n} \sum_{i:y_i<a} y_i \; + \frac{1}{n} \sum_{i:y_i=a} \frac{1}{H} \sum_{k=1}^{H} \; y_{ik}^*$$

The $\lambda' \longleftarrow \lambda$ and we iterate until we find a **fixed point.**

**For the StEM algortihm**, at the iteration $\ell$ we solve the equation (f.o.c. of the maximization step)

$$\frac{\partial}{\partial\theta}\hat{Q}(\theta \mid \theta') = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{H}\sum_{k=1}^{H} \frac{\partial}{\partial\theta} log(f_\theta(\psi(u_{ik}^{\ell} \mid y_i, \theta')) = 0$$

with respect to $\theta$.

So we consider two Simulated EM algorithms : **SimEM and StEM**. When *H* is relatively small, the two simulated EM algorithms may produce different value of $\hat{\theta}_n$.

Table: **Monte Carlo experiment. Exponential duration model with censoring time *a* and parameter $\lambda_0$.**

| | **H=1** | **H=20** | **H=100** | **H=1000** |
|---|---|---|---|---|
| Mean MLE | 0.010004 | 0.0100088 | 0.009998 | 0.010001 |
| Stand. Dev. | 0.0004721 | 0.000477 | 0.000465 | 0.000463 |
| Mean SimEM | 0.009871 | 0.009946 | 0.009987 | 0.009998 |
| Stand. Dev. | 0.000587 | 0.000483 | 0.000465 | 0.000463 |

Note : N=1000. Monte Carlo replications=1000. $a = 60$. $\lambda_0 = 0.01$. $\simeq$ 52% obs censored.

The table on the previous slide regroups the results of a Monte Carlo simulation. The number of MC replications is fixed to 1000. 1000 data sets of size $n = 1000$ are generated independently. For each individual, we draw a duration $y_i^*$ distributed as an exponential random variable with density

$$f_{\theta_0}(y^*) = \lambda_0 \exp(-\lambda_0 y^*)$$

where $\theta_0 = 0.01$.

The observation scheme is such that duration are right censored when $y_i^* \geq a$ (for $a = 60$).

For each data sets, we calculate the MLE and the SimEM estimator. The table, for each value of the number of random uniforms for the SimEM algorithm (H=1,20,100,1000), contains the average and the standard deviation of the ML and SimEM estimates.

Both estimators are in average close to the true value of the parameter. The empirical standard deviations are similar for $H$ large.

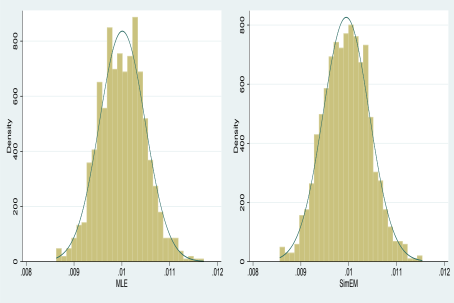

Figure: 1. Distributions of ML and SimEM estimates ($H = 20$)

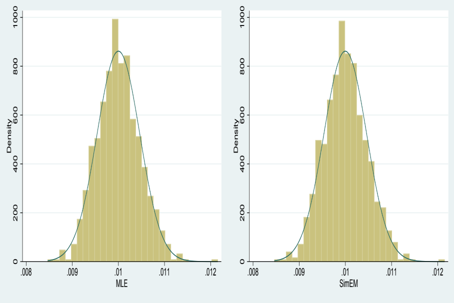

Figure: 2. Distributions of ML and SimEM estimates ($H = 1000$)

Ruud (1991) gives some applications of the use of EM algorithm for econometric models. Let us consider the latent variable

$$Y^*_{ij} = x'_i \beta_j + u_{ij},$$

where i is the index of individual, j the index of equation and $u_i \sim N(0, \Omega)$. $\beta_j$ is a vector of parameters, $x_i$ is a vector of $K$ explanatory variables.

- Multinomial probit : $y_{ij} = 1$ if $y^*_{ij} = \max(y^*_{i1}, \ldots, y^*_{iJ})$, and $y_{ij} = 0$ otherwise.
- Ordered probit : $y_i = g$ when $\alpha_g \leq y^*_i < \alpha_{g+1}$, where $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_G < \alpha_{G+1} = \infty$.
- Non random selection (2): $y_{i1} = \mathbb{1}\{y^*_{i1} > 0\}$ and $y_{i2} = y_{i1} \ y^*_{i2} > 0$ (i.e. $J = 2$).
- Non random selection (1) : $y_i = \mathbb{1}\{y^*_i > 0\} y^*_i$ (i.e. $J = 1$).

We can add, for instance, models with unobserved heterogeneity (panel data models, duration models, consumer choice, see Kenneth Train, etc).

### Simulated EM algorithm in the literature

StEm : Celeux and Diebolt (1985), Diebolt and Celeux (1993), for $H = 1$. Baillif et al. (2021).

SimEM : Ruud (1991), Train (2008)

MCEM : Wei and Tanner (1990), in the E-step, the conditional expectation is calculated via Monte Carlo integration ($H = \infty$). In practice, as $H < \infty$, the results applies : StEM if we draw new uniforms at each iteration or SimEm if we re-use the uniform random draws.

Let $f_\theta(y^*)$, $\theta \in \Theta \subseteq \boldsymbol{R}^p$, denote the density of $Y^*$ and $s_{Y^*}(\theta) = \frac{\partial}{\partial\theta} f_\theta(Y^*)$ the score function. We assume the score function as a zero expectation for $\theta_0$ (model correctly specified). $\theta_0$ is the true value of the vector $\theta$.

The variance covariance matrix $V(\theta) = E_\theta[s_{Y^*}(\theta)\ s_{Y^*}(\theta)']$ (**complete data**). The density of the conditional distribution of $Y^*$ given $Y = y$ is $\kappa_\theta(Y^* \mid y)$. The corresponding score function is denoted $s_{Y^*|y}(\theta)$. Let $I_y(\theta) = E_\theta[s_{Y^*|y}(\theta)\ s_{Y^*|y}(\theta)']$ denote variance of $Y^*$ given $Y = y$.

Let $s_Y(\theta)$ denote the score function of the marginal distribution of $Y$. Let us consider the variance $I(\theta) = E_\theta[s_Y(\theta)\ s_Y(\theta)']$ of $Y$ (**incomplete data**).

**Assumption** : $I(\theta_0)$ and $V(\theta_0)$ are positive definite.

Let us remark that $s_y(\theta) = s_{Y^*}(\theta) - s_{Y^*|y}(\theta)$ and $I(\theta) = V(\theta) - E_\theta[I_Y(\theta)]$.

## Theorem (1)

*Asymptotics, StEM algorithm (Nielsen, 2000)*
*Under regularity assumptions (singularly if the markov chain is erogodic), then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, \Sigma_H(\theta_0))$$

*where*
$\Sigma_H(\theta_0) = I(\theta_0)^{-1} + \frac{1}{H} V(\theta_0)^{-1} E_{\theta_0}[I_Y(\theta_0)] V(\theta_0)^{-1} (I - F(\theta_0)^2)^{-1}$,
$F(\theta_0) = E_{\theta_0}[I_Y(\theta_0)] V(\theta_0)^{-1}$ *and I is the identity matrix.*

$\hat{\theta}_n$ is a random variable drawn from the stationary initial distribution of the Markov chain corresponding to the StEM algorithm. The stationary distribution of the Markov Chain converges to a normal distribution as $n$ becomes large.

Let us consider the derivative of the objective function (M step)

$$G_n(\theta') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{H}\sum_{k=1}^{H}\frac{\partial}{\partial\theta} log(f_\theta(\psi(u_{ik} \mid y_i, \theta'))_{|_{\theta=\theta'}}$$

$G_n(\theta) \xrightarrow{\mathcal{P}} G(\theta)$ (Law of large number). Some technical conditions are needed to obtain uniform convergence.

Assumptions:

(A1) $\sup_{\theta\in C} || G_n(\theta) - G(\theta) || /(1+ || G_n(\theta) || + || G(\theta) ||) \xrightarrow{\mathcal{P}} 0$, for a compact neighbourhood, C, of $\theta_0$.

(A2) $\sup_{||\theta-\theta_0||<\delta_n}(\sqrt{n}||G_n(\theta) - G(\theta) - G_n(\theta_0)||)/(1+\sqrt{n}||G_n(\theta)|| + \sqrt{n}||G(\theta)||) \xrightarrow{\mathcal{P}} 0$, for any $\delta_n \to 0$.

### Theorem (2)

*Asymptotics, SimEM algorithm (Nielsen, 2000)*

*Under (A1) there is exists* $\hat{\theta}_n$*, an asymptotic local minimum of* $||G_n(\theta)||$*, such that* $\hat{\theta}_n \xrightarrow{\mathcal{P}} \theta_0$*.*

*Under (A2), if* $\hat{\theta}_n$ *is consistent, then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, I(\theta_0)^{-1} + \frac{1}{H} I(\theta_0)^{-1} \underset{\theta_0}{E}[I_Y(\theta_0)] I(\theta_0)^{-1})$$

### Theorem (3)

*Comparison of SimEM and StEM algorithms (Nielsen, 2000)*

*Let $H$ denote the number of random numbers we draw per iteration. For the same value of $H$, the estimator we obtain from the StEM algorithm has a smaller asymptotic variance than the asymptotic variance of the estimator resulting from the SimEM algorithm.*

Table: **Monte Carlo experiment. Exponential duration model with censoring time *a* and parameter $\lambda_0$.**

|  | **H=1** | **H=100** | **H=200** |
|---|---|---|---|
| Mean **MLE** | 0.00999111 | 0.00999719 | 0.01000250 |
| Standard Dev | 0.00047180 | 0.00045565 | 0.00047809 |
| Mean **SimEM** | 0.00951538 | 0.01005405 | 0.00998016 |
| Standard Dev | 0.00058965 | 0.00045668 | 0.00047958 |
| Mean **StEM** | 0.00999263 | 0.00999533 | 0.01000003 |
| Standard Dev | 0.00053429 | 0.00045658 | 0.00047919 |

Note : N=1000. Monte Carlo replications=1000. $a = 60$. $\lambda_0 = 0.01$.

Let us consider $n$ individuals and for each individual we observe $T$ realizations of a vector $y_{it}$ of $q$ elements. The realizations are independent across individuals but not (necessarly) across time. Let $y_i = (y_{i1}, \dots, y_{iT})'$.

We can have $K$ different (unobserved) types for individuals. Let $p_k$ denote the probability to belong to type $k$ and $p = (p_1, \dots, p_K)'$.

Let $f_\theta(y \mid k)$ denote the density of $Y$ given the type $k$ and $\theta = (\theta_1, \dots, \theta_E)'$ is a vector of parameters.

The density of $Y_i$ is

$$f_{\theta,p}(y_i) = \sum_{k=1}^{K} p_k \; f_\theta(y_i \mid k)$$

The conditional probability that individual belongs to type $k$ is

$$Prob(k \mid y_i; \theta, p) = \frac{p_k \, f_\theta(y_i \mid k)}{f_{\theta,p}(y_i)} \tag{1}$$

The log-likelihood function is

$$L(\theta, p) = \sum_{i=1}^{n} \log(f_{\theta,p}(y_i))$$

Maximizing the likelihood function under the restriction that $\sum_{k=1}^{K} p_k = 1$ and using equation (1), the maximum likelihood estimator of $p_k$ is

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \tag{2}$$

## Theorem

*The MLE of $\theta$ is a solution of the optimization problem*

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} \sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_\theta(y_i \mid k))$$

Proof:

$\hat{\theta}$ is solution of the optimization problem

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \hat{p}_k f_\theta(y_i \mid k))$$

Then the MLE of $\theta$ is a solution of the first order condition (f.o.c.):

$$\hat{\theta} = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_k \frac{\frac{\partial f_\theta(y_i|k)}{\partial \theta} \big|_{\theta=\hat{\theta}}}{f_{\hat{\theta},\hat{p}}(y_i)} = 0$$

$$\hat{\theta} = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_k \frac{f_\theta(y_i \mid k)}{f_\theta(y_i \mid k)} \frac{\frac{\partial f_\theta(y_i|k)}{\partial \theta} \,|_{\theta=\hat{\theta}}}{f_{\hat{\theta},\hat{p}}(y_i)} = 0$$

$$\hat{\theta} = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{p}_k \frac{f_\theta(y_i \mid k)}{f_{\hat{\theta},\hat{p}}(y_i)} \frac{\frac{\partial f_\theta(y_i|k)}{\partial \theta} \,|_{\theta=\hat{\theta}}}{f_\theta(y_i \mid k)} = 0$$

The is a condition that can be obtained if we consider the following optimization program

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\hat{p}_k f_{\hat{\theta}}(y_i \mid k)}{\sum_{k=1}^{K} \hat{p}_k f_{\hat{\theta}}(y_i \mid k)} \log(f_\theta(y_i \mid k))$$

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} \sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_\theta(y_i \mid k))$$

QED

Then the MLE $\hat{\theta}$ of $\theta$ is a solution of

$$\sum_{i=1}^{n}\sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \frac{\partial \log(f_\theta(y_i \mid k))}{\partial \theta} \mid_{\theta=\hat{\theta}} = 0 \qquad (3)$$

Consequently,

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n}\sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_\theta(y_i \mid k)) \qquad (4)$$

The MLE $\hat{\theta}$ maximizes

- the log of the likelihood function $\sum_{i=1}^{n} \log(\sum_{k=1}^{K} \hat{p}_k \; f_\theta(y_i \mid k))$
- the conditional expectation (indeed an component of a likelihood function) $\sum_{i=1}^{n} \sum_{k=1}^{K} \; Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_\theta(y_i \mid k))$

where $\hat{\theta}$ and $\hat{p}$ are the MLE if $\theta$ and $p$ respectively. The E-step consists to maximize the last object with respect to $\theta$.

**EM algorithm**:

- At the iteration $\ell$, $\theta^0$ is replaced by $\theta^{\ell-1}$ and $p^0$ is replaced by $p^{\ell-1}$.
- The E-step consists to calculate $Prob(k \mid y_i; \theta^{\ell-1}, p^{\ell-1})$.
- The M-step consists to use equation (2) to obtain $p^\ell$ and the equation (4) to deduce $\theta^\ell$.

The algorithm is such that we iterate until we achieve convergence.

### Theorem

*MLE of $p_k$*

$$\hat{p}_k = \frac{1}{n}\sum_{i=1}^{n} Prob(k \mid y_i; \hat{\theta}, \hat{p})$$

where $Prob(k \mid y_i; \theta, p) = \frac{p_k\, f_\theta(y_i|k)}{f_{\theta,p}(y_i)}$.

Proof: $L(\theta, p) = \sum_{i=1}^{n} ln(f_{\theta,p}(y_i)) = \sum_{i=1}^{n} ln(\sum_{k=1}^{K} p_k\, f_\theta(y_i \mid k))$ is the likelihood function we maximize under the constraint that $\sum_{k=1}^{K} p_k = 1$.

$$\frac{\partial}{\partial p_k} L(\theta, p) = \sum_{i=1}^{n} \frac{f_\theta(y_i \mid k) - f_\theta(y_i \mid K)}{f_{\theta,p}(y_i)} = 0$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{p_k f_\theta(y_i \mid k)}{f_{\theta,p}(y_i)} = \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{p_k f_\theta(y_i \mid K)}{f_{\theta,p}(y_i)} = \sum_{i=1}^{n} \frac{f_\theta(y_i \mid K)}{f_{\theta,p}(y_i)}$$

Then

$$n = \sum_{i=1}^{n} \frac{f_\theta(y_i \mid K)}{f_{\theta,p}(y_i)}$$

as $\sum_{k=1}^{K} p_k f_\theta(y_i \mid k) = f_{\theta,p}(y_i)$ and

$$\sum_{i=1}^{n} \frac{f_\theta(y_i \mid k)}{f_{\theta,p}(y_i)} = n$$

$$\sum_{i=1}^{n} \frac{p_k f_\theta(y_i \mid k)}{f_{\theta,p}(y_i)} = p_k n$$

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}_k f_{\hat{\theta}}(y_i \mid k)}{f_{\hat{\theta},\hat{p}}(y_i)} = \frac{1}{n} \sum_{i=1}^{n} Prob[k \mid y_i; \hat{\theta}, \hat{p}]$$

Q.E.D.

Let us consider that $\theta = (\theta_1, \theta_2)'$. We can maximize the equation (4) (M-step) jointly with respect to $\theta$ or by maximizing the expression with respect to $\theta_2$ taking the last value of $\hat{\theta}_1$ as fixed and then, maximizing the expression with respect to $\theta_1$ given the last value $\hat{\theta}_2$ of $\theta_2$.

The sequential way of resolution of the EM algorithm is called **Expectation-Conditional Maximization (ECM)**, see Meng and Rubin (1993) and Arcidiacono and Bailey Jones (2003). The **ECM algorithm** keeps the convergence properties of the EM algorithm.

Singularly, let us assume that

$$f_{\theta_1,\theta_2}(y_i \mid k) = f_{1,\theta_1}(y_i \mid k)\; f_{2,\theta_1,\theta_2}(y_i \mid k)$$

The log-likelihood function is

$$L(\theta, p) = \sum_{i=1}^{n} \log(\sum_{k=1}^{K} p_k \; f_{1,\theta_1}(y_i \mid k) \; f_{2,\theta_1,\theta_2}(y_i \mid k))$$

The expression **cannot** directly **be factorized** as a product $L_1(\Theta_1)L(\Theta_1, \Theta_2)$ in oder to implement a sequential estimation. Within the EM algorithm, equation (4) can be (re)written

$$(\hat{\theta}_1, \hat{\theta}_2) = \underset{\theta_1,\theta_2}{argmax} \sum_{i=1}^{n} \sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_{1,\theta_1}(y_i \mid k))$$

$$+ \sum_{i=1}^{n} \sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_{2,\theta_1,\theta_2}(y_i \mid k))$$

We can resolve sequentially

$$\hat{\theta}_1 = \underset{\theta_1}{argmax} \sum_{i=1}^{n} \sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_{1,\theta_1}(y_i \mid k)) \quad (5)$$

$$\hat{\theta}_2 = \underset{\theta_2}{argmax} \sum_{i=1}^{n} \sum_{k=1}^{K} Prob(k \mid y_i; \hat{\theta}, \hat{p}) \log(f_{2,\hat{\theta}_1,\theta_2}(y_i \mid k)) \quad (6)$$

The EM algorithm allows to estimate $\theta$ sequentially in such a way that, for each stage, we use the estimates of other parameters obtained from the previous stage (each step comprises two stages).

We can remark that we do not need to derivate equation $f_{2,\hat{\theta}_1,\theta_2}(y_i \mid k)$ with respect to $\theta_1$ so that the estimator is **less efficient that the FIML** one (Acidiacono and Jones, 2003). Moreover, the EM algorithm allows to obtain the FIML estimates (**estimates** obtained using *ECM* **algorithm** $\neq$ **FIML estimates**).

When we use a star, the corresponding parameters are the population parameters. **At the level of population**

$$(\theta^*, p^*) = \underset{\theta,p}{argmax}\ \underset{y,k}{E} \log(p_k\ f_{1,\theta_1}(y \mid k)\ f_{2,\theta_1,\theta_2}(y \mid k)) \qquad (7)$$

where $E_{y,k}$ states the expectation with respect to the distribution of the $(y, k)$.

The expectation with respect to the joint distribution of $(y, k)$ is also the expectation with respect to the distribution of $y$ of the conditional expectation with respect to $k$ given $y$.

Consequently,

$$(\theta^*, p^*) = \underset{\theta,p}{argmax} \underset{y}{E} \left[ \sum_{k=1}^{K} Prob[k \mid y; \theta^*, p^*] \right. \\ \left. \times \log(p_k f_{1,\theta_1}(y \mid k)\ f_{2,\theta_1,\theta_2}(y \mid k)) \right] \tag{8}$$

where $Prob[k \mid y; \theta^*, p^*] = \frac{p_k^* f_{\theta^*}(y|k)}{f_{\theta^*,p^*}(y)}$ and $E_y$ is the marginal expectation with respect to the distribution of $y$.
From the previous equation we can remark that

$$\theta_2^* = \underset{\theta_2^*}{argmax} \underset{y}{E} \left[ \sum_{k=1}^{K} Prob[k \mid y; \theta^*, p^*]\ \log(f_{2,\theta_1^*,\theta_2}(y \mid k)) \right]$$

It is the population analog of the equation (6).

## Asymptotic behavior of the Sequential estimator

Let us remark that $f_{1,\theta_1}(y \mid k)$ is a conditional density and let us consider the population of individuals with the same value of $y$. Each individual in this population has a probability $Prob[k \mid y; \theta^*, p^*]$ to belong to type $k$.
The population objective function is

$$\underset{y}{E}\left[\sum_{k=1}^{K} Prob[k \mid y; \theta^*, p^*] \; \log(f_{1,\theta_1}(y \mid k))\right] +$$

$$\underset{y}{E}\left[\sum_{k=1}^{K} Prob[k \mid y; \theta^*, p^*] \; \log(Prob[k \mid y; \theta^*, p^*)\right]$$

that have to be maximized with respect to $\theta_1$.
The first order conditions are

$$\underset{y}{E}\left[\sum_{k=1}^{K} Prob[k \mid y; \theta^*, p^*] \; \frac{\partial \log(f_{1,\theta_1^*}(y \mid k))}{\partial \theta_1}\right] = 0$$

The population moment conditions are

$$\underset{y}{E}\begin{bmatrix} \sum_{k=1}^{K} Prob[k \mid y;\theta^*,p^*]\, \frac{\partial \log(f_{1,\theta_1^*}(y|k))}{\partial\theta_1} \\ \sum_{k=1}^{K} Prob[k \mid y;\theta^*,p^*]\, \frac{\partial \log(f_{2,\theta_1^*,\theta_2^*}(y|k))}{\partial\theta_2} \\ Prob[1 \mid y;\theta^*,p^*] - p_1^* \\ \vdots \\ Prob[K \mid y;\theta^*,p^*] - p_K^* \end{bmatrix} = 0 \qquad (9)$$

where $Prob[k \mid y;\theta^*,p^*] = \frac{p_k^* f_{\theta^*}(y|k)}{f_{\theta^*,p^*}(y)}$ and $E_y$ is the expectation with respect to the marginal distribution of $y$.

Under the usual regularity conditions (Hansen, 1982 or Newey and MCFadden, 1994), $(\hat{\theta}_1, \hat{\theta}_2, \hat{p})'$ solution of (2), (5) and (6) are **consistent and asymptotically normal.**

The variance-covariance matrix of this method of moments estimator is given by the standard formula. The calculation of this variance-covariance matrix involve all these **moment conditions** despite that $p$, $\theta_1$ and $\theta_2$ **are estimated sequentially**.

If we consider equation (9) we can remark that the two first components of the vector of moment conditions are not part of the score vector of the MLE since

$$\frac{\partial \log(f_{\theta^*,p^*}(y))}{\partial\theta_1} = \sum_{k=1}^{K} \frac{p_k^* f_{\theta^*}(y|k)}{f_{\theta^*,p^*}(y)} \left[ \frac{\partial \log(f_{1,\theta_1^*}(y|k))}{\partial\theta_1} + \frac{\partial \log(f_{2,\theta_1^*,\theta_2^*}(y|k))}{\partial\theta_1} \right]$$

$$\frac{\partial \log(f_{\theta^*,p^*}(y))}{\partial\theta_2} = \sum_{k=1}^{K} \frac{p_k^* f_{\theta^*}(y|k)}{f_{\theta^*,p^*}(y)} \frac{\partial \log(f_{2,\theta_1^*,\theta_2^*}(y|k))}{\partial\theta_2}$$

So the **sequential estimator is not as efficient as the MLE.**

In order to satisfy the regularity conditions usually assumed in order to obtain **consistency and asymptotic normality** (see Newey and McFadden, 1994), Acidiacono and Jones (2003) assume that the expectation of the log-likelihood function has a unique maximum for $(\theta^*, p^*)$ and the moment conditions (9) are the features of this optimum.

Let $Q(\theta, p)$ denote the negative of the GMM criterium function behind the sequential estimator. The mixture of conditional likelihood is **not in general globally concave.** The objective function on the data set can have multiple local maximizers. The approach consists to select among these local maximizers, the one that maximizes the log-likelihood (Wu, 1983).

Consequently, we should use this likelihood criterion when the system (9) has more than one solution. This rule of selection ensure **consistency of the sequential estimator** (ECM estimator). The ECM estimator is distinct from the FIML estimator.

Indeed, Bonhomme (2006) gives an alternative estimate of the variance-covariance matrix of the parameters.

Assumption : There exists a partition $\theta = (\theta_1, \theta_2)$ such that for all $\theta_1$, $\theta_2$, $\theta'_1$ and $k = 1, \dots, K$ we have

$$f_\theta(y_i \mid k) = f_{1,\theta_1}(y_i \mid k) f_{2,\theta_1,\theta_2}(y_i \mid k), \tag{10}$$

$$\int f_{1,\theta_1}(y \mid k)\ f_{2,\theta'_1,\theta_2}(y \mid k)\ d\,y = 1. \tag{11}$$

In order to apply the GMM theory, let us consider

$$\Psi(y_i;\xi) = \begin{pmatrix} Prob[k = 1 \mid y_i;\xi] - p_1 \\ \dots \\ Prob[k = K \mid y_i;\xi] - p_K \\ \sum_{k=1}^K Prob[k \mid y_i;\xi]\ \frac{\partial \ln(\ f_{1,\theta_1}(y_i|k)\ )}{\partial\theta_1} \\ \sum_{k=1}^K Prob[k \mid y_i;\xi]\ \frac{\partial \ln(\ f_{2,\theta_1,\theta_2}(y_i|k)\ )}{\partial\theta_2} \end{pmatrix}$$

where $\xi = (\theta_1, \theta_2, p)$.

Then the moment conditions are

$$E[\Psi(y_i;\xi^0)] = 0$$

The variance-covariance matrix of the estimator is given by the sandwich formula (see Newey and McFadden, 1994)

$$\Sigma = E[\frac{\partial\Psi(y_i;\xi^0)}{\partial\xi'}]^{-1}\ E[\Psi(y_i;\xi^0)\Psi(y_i;\xi^0)']\ E[\frac{\partial\Psi(y_i;\xi^0)}{\partial\xi'}]^{-1}$$

### Theorem

*For all* $\xi = (\theta_1, \theta_2, p)'$ *we have*

$$\int \Psi(y_i;\xi)\ f(y;\xi)\ d\,y = 0 \tag{12}$$

*where* $f(y;\xi) = \sum_{k=1}^{K} p_k\ f_\theta(y \mid k)$.

## Theorem

*The variance covariance matrix* $\Sigma$ *depends on the first derivatives of the likelihood and on the posterior probabilities only.*

$$E\left[\frac{\partial\Psi(y;\xi^0)}{\partial\xi'}\right] = -E\left[\Psi(y_i;\xi^0)\ \frac{\partial \ln(\ f(y;\xi^0)\ )}{\partial\xi'}\right] \tag{13}$$

For the expression of $\frac{\partial \ln(\ f(y;\xi^0)\ )}{\partial\xi'}$ see the Appendix.

This part is dedicated to the **extension of the sequential approach** to general moment conditions in models incorporating unobserved heterogeneity (we consider a finite mixture but it can be extended to general mixtures).

Any function $h_\theta(y, k)$ that verifies

$$\underset{y,k}{E}\,[h_{\theta^*}(y, k)] = 0$$

also verifies

$$\underset{y}{E}[\sum_{k=1}^{K} Prob[k \mid y; \theta^*, p^*] h_{\theta^*}(y, k)] = 0$$

This **population condition** is similar to the **sample condition** (3)

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Prob[k \mid y; \hat{\theta}, \hat{p}] \frac{\partial \log(f_{\hat{\theta}}(y_i \mid k))}{\partial\theta} = 0$$

where $\hat{\theta}$ is the ML estimator of $\theta$.

In this case we have

$$h_\theta(y,k) = \frac{\partial \log(f_{\hat{\theta}}(y_i \mid k))}{\partial \theta}$$

**Example**: Switching regressions (Kiefer, 1980).

Let us consider the following linear regression model

$$y_i = x_i' \beta_k^* + u_i$$

where $x_i$ is a vector of explanatory variables for individual $i$, $\beta_k$ is a vector of parameters and $u_i$ is a normal error term (0 expectation and variance $\sigma^2$). $u_i$ is assumed to be independent of the unobserved type $k$ and $x_i$. Let $\theta = (\beta_1', \ldots, \beta_k')'$. Let us remark that

$$\underset{y,x}{E} \left( Prob[k \mid y, x; \theta^*, p^*] \; x(y - x'\beta_k^*) \right) = 0$$

for all $k \in \{1, \ldots, K\}$

## Generalizations of the Sequential estimator (ECM)

Let us consider a diagonal weighting matrix $\hat{W}_k$ such that the element $i$ of the main diagonal is $\sqrt{Prob[k \mid y_i, x_i; \hat{\theta}, \hat{p}]}$.

The algorithm consists to estimate $\beta_k$ using matrices $\hat{W}_k X$ and vectors $\hat{W}_k y$. Then, we use these estimates in order to update obtain a new value of $\hat{W}_k$.

In order to update $\hat{W}_k$, we need

$$Prob[k \mid y_i, x_i; \theta, p; \sigma^2] = \frac{p_k \; f_{\beta_k,\sigma^2}(y_i \mid x_i; k)}{f(y_i \mid x_i; \theta, p, \sigma^2)}$$

where

$$f(y_i \mid x_i; \theta, p, \sigma^2) = \sum_{k=1}^{K} p_k \; f_{\beta_k,\sigma^2}(y_i \mid x_i; k)$$

and

$$p_k = \frac{1}{n} \sum_{i=1}^{n} Prob[k \mid y_i, x_i; \theta, p; \sigma^2]$$

Let us consider the case such that we have two regimes ($K = 2$).

In this case the distribution of $k \mid x; y; \theta, p, \sigma^2$ is logistic. Let us assume that $\delta_i = 1$ if the individual $i$ is in regime 1 and $\delta_i = 0$ otherwise.

The conditional probability that $\delta_i$ is equal to 1 is

$$Prob[k \mid y_i, x_i; \theta, p; \sigma^2] = \frac{1}{1 + \exp(a_i + b_i\, y_i)} \equiv \omega_i \tag{14}$$

where $a_i = \ln(\frac{1-p_1}{p_1}) + \frac{1}{2\sigma^2}((x_i'\beta_1)^2 - (x_i'\beta_2)^2)$ and $b_i = \frac{1}{\sigma^2}(x_i'\beta_2 - x_i\beta_1)$.

Let us consider the matrix $W = diag(\omega_1, \dots, \omega_n)$, where $n$ is the number of individuals in the sample.

Let us consider the following estimators

$$\hat{\beta}_1 = (X'WX)^{-1}X'Wy \tag{15}$$

## Generalizations of the Sequential estimator (ECM)

$$\hat{\beta}_2 = (X'(I_n - W)X)^{-1} X'(I_n - W)y \tag{16}$$

$$\begin{aligned}\hat{\sigma}^2 = \quad & \tfrac{1}{n}[\ (y - X\beta_1)' W((y - X\beta_1) \\ & \quad +(y - X\beta_2)'(I_n - W)((y - X\beta_2)\ ]\end{aligned} \tag{17}$$

$$\hat{p}_1 = \frac{1}{n} trW \tag{18}$$

**Algorithm**: Start with an initial value for $\xi = (\beta_1, \beta_2, \sigma^2, p_1)'$, namely $\xi^{(1)}$. The index $k$ is set to the value 1.

1) Calculate the corresponding $\omega_i$ using (14) (namely $\omega_i^{(k)}$).

2) $k \leftarrow k + 1$.Update the value of $\xi$ using (15)-(18) to obtain $\xi^{(k)}$.

3) Then go to 1) or stop if the convergence is achieved.

The algorithm allows to obtain the **MLE** of $\xi$ (see Kiefer 1980).

Remark : In the paper of Baillif et al. (2021), the algorithm exploit similar relationship to system (15)-(18). The main difference is that the "regimes" are not observed as there are discrete unobserved heterogeneity components.

Consequently, they draw the "regimes" - unobserved heterogeneity components - conditionally to current values of parameters - at step k - in a multinomial distribution.

Remark : The problem of estimation of models with discrete unobserved heterogeneity components is important in econometrics (see Heckman and Singer, 1984, Bonhomme, Lamadon and Manresa, 2017).

## Generalizations of the Sequential estimator (ECM)

Some models with finite mixtures:

Baillif, de Lapparent and Kazagli (2021): Study of real estate Price.

Bonhomme and Robin (2006): Wage and Employment dynamics.

Cameron and Heckman (1998, 2001): Impact if family background on educational attainment.

Eckstein and Wolpin (1999) : Models of dynamic discrete choice.

Heckman and Singer (1984): Models for duration data.

Some models with finite mixtures (continued):

Keane and Wolpin (1997) : Labor economics.

Kiefer (1980): Switching regressions.

Mroz (1999) : Impact of a Dummy Endogenous Variable on a Continuous Outcome.

Train (2008) : Mixed logit model of households' choices among alternative-fueled vehicle.

Provencher, B., K. Baerenklau and R. Bishop (2002) : Model of recreational angling.

Arcidiacono, P., Jones, J.B., 2003. Finite mixture distributions, sequential likelihood and the EM algorithm. Econometrica, 71(3), 933-946.

Baillif, M. , de Lapparent, M., Kazagli, E., 2021. A Hybrid Approach to Real Estate Price Definition: A Case Study in Western Switzerland. Revue Economique, 72(6), 1055–1077.

Bonhomme, S., Robin, J.-M., 20006. Modeling Individual Earnings Trajectories Using Copulas: France, 1990-2002. In Structural Models of Wage and Employment Dynamics. Henning Bunzel, Bent Christensen, George R. Neumann and Jean-Marc Robin eds. Elsevier.

Bonhomme, S., 2006, Standard Errors Estimation in Mixture of Partial Likelihood Models. University of Chicago.

Bonhomme, S., Lamadon, T., Manresa, E., 2017, Discretizing Unobserved Heterogeneity. To be published in Econometrica.

Cameron, S., Heckman, J.J., 1998. Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males. Journal of Political Economy, 106, 262-333.

Cameron, S., Heckman, J.J., 2001. The Dynamics of Educational Attainment for Black, Hispanic, and White Males. Journal of Political Economy, 109, 455-499.

Celeux, G., Diebolt, J., 1985. The SEM algorithm: a probabilistic teacher algorithm derived from EM algorithm for the mixture problem. Computational Statistics Quarterly, 2, 73-82.

Chung, Y., Lindsay, B.G., 2015. Convergence of the EM algorithm for continuous mixing distributions. Statistics and Probability Letters. 96, 190-195.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B39(1), 1-38.

Diebolt, J., Celeux, G., 1993. Asymptotic properties of a stochastic EM algorithm for estimating proportions. Communications Statistics and Stochastic Models, 9, 599-613.

Eckstein, Z., Wolpin, K., 1999. Why Youths Drop Out of High School: The Impact of Preferences, Opportunities and Abilities. Econometrica, 67, 1295-1339.

Hansen, L., 1982. Large Sample Properties of Generalized Method of Moments Estimators. Econometrica, 50, 1029-1054.

Heckman, J.J., Singer, B., 1984. A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. Econometrica, 52(2), 271–320.

Keane, M., Wolpin, K., 1997. The Career Decisions of Young Men. Journal of Political Economy, 105, 473-522.

Kiefer, N., 1980. A Note on Switching Regressions and Logistic Discrimination. Econometrica, 48(4), 1065–1068.

Nielsen, S.F., 2000. On simulated EM algorithms. Journal of Econometrics, 96, 267-292.

Nielsen, S.F., 2000. The stochastic EM algorithm: estimation and asymptotic results. Bernoulli, Bernoulli 6(3), 2000, 457-489.

Meng, X., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika, 80, 267-278.

Mroz, T. A., 1999. Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome. Journal of Econometrics, 92, 233-274.

Newey, W. K., McFadden, D. 1994. Large sample estimation and hypothesis testing. In Handbook of Econometrics, Volume 4, ed. by R. F Engle and D. L. McFadden. Amsterdam: North Holland, 2113-2245.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica, 57, 1027-1057.

Provencher, B., Baerenklau, K., Bishop, R., 2002. A finite mixture logit model of recreational angling with serially correlated random utility. American Journal of Agricultural Economics 84 (4), 1066-1075.

Ruud, P.A., 1991, Extensions of estimation methods using the EM algorithm. Journal of Econometrics, 49, 305-341.

Train, K. E., 2008. EM Algorithms for nonparametric estimation of mixing distributions, Journal of Choice Modelling, 1(1), 2008, 40-69.

Wei, G.C.G., Tamer, 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, 699-704.

Wu, C.F.J.,1983. On the convergence properties of the EM algorithm. Ann. Statist. 11(1), 95-103.

The expression of $\frac{\partial \ln( f(y;\xi^0) )}{\partial \xi'}$, where $\xi = (\theta_1, \theta_2, p)'$ (see sequential likelihood section).

$$\frac{\partial \ln( f(y;\xi^0) )}{\partial p_k} = \frac{Prob[k \mid y;\xi^0]}{\partial p_k}, k = 1, \dots, K, \tag{19}$$

$$\frac{\partial \ln( f(y;\xi^0) )}{\partial \theta_1} = \sum_{k=1}^{K} Prob[k|y;\xi^0] \frac{\partial(\ln( f_{1,\theta_1^0}(y|k) ) + \ln( f_{2,\theta_1^0,\theta_2^0}(y|k) ))}{\partial \theta_1} \tag{20}$$

$$\frac{\partial \ln( f(y;\xi^0) )}{\partial \theta_2} = \sum_{k=1}^{K} Prob[k \mid y;\xi^0] \frac{\partial(\ln( f_{2,\theta_1^0,\theta_2^0}(y \mid k) ))}{\partial \theta_2} \tag{21}$$