KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Fitting mixtures of Erlangs to uncensored and untruncated data using the EM algorithm - Addendum

Verbelen R.

KBI_1401B

Fitting mixtures of Erlangs to uncensored and untruncated data using the EM algorithm

Roel Verbelen

October 23, 2013

Abstract

In this addendum, we present the EM algorithm of Lee and Lin (2010) custimized for fitting mixtures of Erlang distributions with a common scale parameter to uncensored and untruncated data. We work out the details with zero-one component indicators inspired by McLachlan and Peel (2001) and Lee and Scott (2012).

1 Likelihood

Let $\boldsymbol{x} = (x_1, \dots, x_n)$ be an observed sample from the mixture of Erlang distributions with density given by

$$f(x; \boldsymbol{\alpha}, \boldsymbol{r}, \theta) = \sum_{j=1}^{M} \alpha_j \frac{x^{r_j-1} e^{-x/\theta}}{\theta^{r_j}(r_j-1)!} = \sum_{j=1}^{M} \alpha_j f(x; r_j, \theta) \qquad \text{for } x \geqslant 0\,. \tag{1}$$

The parameters to be estimated are the mixing proportions or weights $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$ with $\alpha_j > 0$ and $\sum_{j=1}^{M} \alpha_j = 1$ and the common scale parameter θ, which are bundled by denoting $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$. The number of Erlangs M in the mixture and the corresponding positive integer shapes $\boldsymbol{r}$ are fixed. The value of M is, in most applications, however unknown and has to be inferred from the available data, along with the shape parameters. The log likelihood is given by

$$l(\boldsymbol{\Theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{M} \alpha_j \frac{x_i^{r_j-1} e^{-x_i/\theta}}{\theta^{r_j}(r_j-1)!} \right)\,.$$

which is difficult to numerically optimize due to logarithm of a summation.

2 Construction of the complete data vector

The EM algorithm provides a computationally much easier way for fitting this finite mixture. The main clue is to regard the observed sample $\boldsymbol{x} = (x_1, \dots, x_n)$ as being incomplete since their associated component-indicator vectors $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ with

$$z_{ij} = \begin{cases} 1 & \text{if observation } x_i \text{ comes from } j\text{th component density } f(x; r_j, \theta) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, M$, are not available (McLachlan and Peel (2001)). The component-label vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are taken to be realized values of the random vectors

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{\text{i.i.d.}}{\sim} \text{Mult}_M(1, \boldsymbol{\alpha}).$$

The log likelihood of the complete data vector $(\boldsymbol{x}, \boldsymbol{z})$ equals

$$l(\boldsymbol{\Phi}; \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \ln \left(\alpha_j f_X(x_i; r_j, \theta) \right) . \tag{3}$$

The EM algorithm exploits the simpler form of the complete data log likelihood to compute the maximum likelihood estimators based on the observed data.

3 Initial step

Initialization of θ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ is based on the denseness property (see Appendix A):

$$\theta^{(0)} = \frac{\max(\boldsymbol{x})}{r_M} \quad \text{and} \quad \alpha_j^{(0)} = \frac{\sum_{i=1}^{n} I\left(r_{j-1}\theta^{(0)} < x_i \leqslant r_j \theta^{(0)} \right)}{n}, \quad \text{for } j = 1, \ldots, M, \tag{4}$$

with $r_0 = 0$ for notational convenience. These starting values ensure that the initial guess is immediately quite decent.

4 E-step

In the E-step, we take the conditional expectation of the complete log likelihood (3) given the observed data $\boldsymbol{x}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ for $\boldsymbol{\Theta}$. Define, for $i = 1, \ldots, n$ and $j = 1, \ldots, M$, the posterior probability $z_{ij}^{(k)}$ that observation i belongs to the jth component in the mixture,

$$z_{ij}^{(k)} = E(Z_{ij} \mid \boldsymbol{x}; \boldsymbol{\Theta}^{(k-1)}) = \frac{\alpha_j^{(k-1)} f(x_i; r_j, \theta^{(k-1)})}{\sum_{m=1}^{M} \alpha_m^{(k-1)} f(x_i; r_m, \theta^{(k-1)})} . \tag{5}$$

Then

$$\begin{aligned} Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) &= E(l(\boldsymbol{\Theta}; \boldsymbol{x}, \boldsymbol{Z}) \mid \boldsymbol{x}; \boldsymbol{\Theta}^{(k-1)}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} E(Z_{ij} \mid \boldsymbol{x}; \boldsymbol{\Theta}^{(k-1)}) \ln \left(\alpha_j f_X(x; r_j, \theta) \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij}^{(k)} \Big[\ln(\alpha_j) + (r_j - 1) \ln(x_i) - \frac{x_i}{\theta} \\ &\qquad - r_j \ln(\theta) - \ln((r_j - 1)!) \Big] , \end{aligned} \tag{6}$$

The E-step hence reduces to calculating all posterior probabilities.

5 M-step

The M-step requires the global maximization of (6) obtained in the E-step with respect to $\mathbf{\Theta} = (\boldsymbol{\alpha}, \theta)$ with $\alpha_i > 0$, $\sum_{i=1}^M \alpha_i = 1$ and $\theta > 0$. We first maximize (6) with respect to the mixing proportions $\boldsymbol{\alpha}$. This can be done separately of the updated estimate for θ as it requires the maximization of

$$\sum_{i=1}^n \sum_{j=1}^M z_{ij}^{(k)} \ln(\alpha_j) = \sum_{i=1}^n \sum_{j=1}^{M-1} z_{ij}^{(k)} \ln(\alpha_j) + \sum_{i=1}^n z_{iM}^{(k)} \ln\left(1 - \sum_{j=1}^{M-1} \alpha_j\right)$$

with respect to $\alpha_1, \ldots, \alpha_{M-1}$. Note that we implement the restriction $\sum_{j=1}^M \alpha_j = 1$ by setting $\alpha_M = 1 - \sum_{j=1}^{M-1} \alpha_j$. Setting the partial derivatives at $\boldsymbol{\alpha}^{(k)}$ equal to zero yields

$$\left.\frac{\partial Q(\mathbf{\Theta}; \mathbf{\Theta}^{(k-1)})}{\partial \alpha_j}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}} = \sum_{i=1}^n \frac{z_{ij}^{(k)}}{\alpha_j} - \left.\sum_{i=1}^n \frac{z_{iM}^{(k)}}{\alpha_M}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}} = 0 \qquad \text{for } j = 1, \ldots, M-1\,.$$

This implies that the optimizer satisfies

$$\alpha_j^{(k)} = \frac{\sum_{i=1}^n z_{ij}^{(k)}}{\sum_{i=1}^n z_{iM}^{(k)}} \alpha_M^{(k)} \qquad \text{for } j = 1, \ldots, M-1\,. \tag{7}$$

Using the restriction that the mixing weights must sum to one, we obtain

$$1 = \sum_{j=1}^M \alpha_j^{(k)} = \frac{\sum_{i=1}^n \left(\sum_{j=1}^M z_{ij}^{(k)}\right) \alpha_M^{(k)}}{\sum_{i=1}^n z_{iM}^{(k)}} = \frac{n\alpha_M^{(k)}}{\sum_{i=1}^n z_{iM}^{(k)}}\,.$$

Hence

$$\alpha_M^{(k)} = \frac{\sum_{i=1}^n z_{iM}^{(k)}}{n}$$

and by plugging this expression in (7), the same form also follows for $j = 1, \ldots, M-1$:

$$\alpha_j^{(k)} = \frac{\sum_{i=1}^n z_{ij}^{(k)}}{n} \qquad \text{for } j = 1, \ldots, M\,.$$

This solution has a nice intuitive interpretation. The new estimate for the prior probability α_j is the average over all observations i of the posterior probability $z_{ij}^{(k)}$ of belonging to the jth component in the mixture. The optimizer indeed corresponds to a maximum since

$$\begin{aligned}\left.\frac{\partial^2 Q(\mathbf{\Theta}; \mathbf{\Theta}^{(k-1)})}{\partial \alpha_j^2}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}} &= -\sum_{i=1}^n \frac{z_{ij}^{(k)}}{\alpha_j^2} - \left.\sum_{i=1}^n \frac{z_{iM}^{(k)}}{\alpha_M^2}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}} \\ &= -\frac{n^2}{\sum_{i=1}^n z_{ij}^{(k)}} - \frac{n^2}{\sum_{i=1}^n z_{iM}^{(k)}}\end{aligned}$$

for $j = 1, \ldots, M$ and

$$\begin{aligned}\left.\frac{\partial^2 Q(\mathbf{\Theta}; \mathbf{\Theta}^{(k-1)})}{\partial \alpha_j \partial \alpha_m}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}} &= -\left.\sum_{i=1}^n \frac{z_{iM}^{(k)}}{\alpha_M^2}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^{(k)}} \\ &= -\frac{n^2}{\sum_{i=1}^n z_{iM}^{(k)}}\,,\end{aligned}$$

for $j = 1, \dots, M$ and $m = 1, \dots, M$, implying that the matrix of second order partial derivatives is negative definite matrix with a compound symmetry structure.

We next maximize (6) with respect to θ:

$$\begin{aligned} \left.\frac{\partial Q(\mathbf{\Theta};\mathbf{\Theta}^{(k-1)})}{\partial \theta}\right|_{\theta=\theta^{(k)}} &= \left.\sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}^{(k)}\left(\frac{x_i}{\theta^2}-\frac{r_j}{\theta}\right)\right|_{\theta=\theta^{(k)}} \\ &= \left.\frac{1}{\theta^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{M} z_{ij}^{(k)}\right)x_i - \frac{n}{\theta}\sum_{j=1}^{M}\left(\frac{\sum_{i=1}^{n} z_{ij}^{(k)}}{n}\right)r_j\right|_{\theta=\theta^{(k)}} \\ &= \left.\frac{1}{\theta^2}\sum_{i=1}^{n} x_i - \frac{n}{\theta}\sum_{j=1}^{M}\alpha_j^{(k)} r_j\right|_{\theta=\theta^{(k)}} = 0\,. \end{aligned}$$

Hence

$$\theta^{(k)} = \frac{\sum_{i=1}^{n} x_i/n}{\sum_{j=1}^{M}\alpha_j^{(k)} r_j}\,, \tag{8}$$

which is a maximum since

$$\begin{aligned} \left.\frac{\partial^2 Q(\mathbf{\Theta};\mathbf{\Theta}^{(k-1)})}{\partial \theta^2}\right|_{\theta=\theta^{(k)}} &= \left.\frac{-2}{\theta^3}\sum_{i=1}^{n} x_i + \frac{n}{\theta^2}\sum_{j=1}^{M}\alpha_j^{(k)} r_j\right|_{\theta=\theta^{(k)}} \\ &= \left. n\sum_{j=1}^{M}\alpha_j^{(k)} r_j\left[\frac{-2\sum_{i=1}^{n} x_i/n}{\theta^3\sum_{j=1}^{M}\alpha_j^{(k)} r_j}+\frac{1}{\theta^2}\right]\right|_{\theta=\theta^{(k)}} \\ &= n\sum_{j=1}^{M}\alpha_j^{(k)} r_j\left[\frac{-1}{\left(\theta^{(k)}\right)^2}\right] < 0\,. \end{aligned}$$

The new estimate $\theta^{(k)}$ in (8) for the common scale parameter θ equals the sample mean divided by the weighted average shape parameter in the mixture. The updating scheme (8) for the scale parameter makes intuitively sense since the expected value of a mixture of Erlangs equals $E(X) = \sum_{j=1}^{M}\alpha_j r_j \theta$.

The E- and M-steps are iterated until the difference in log likelihood values $l(\mathbf{\Theta}^{(k)};\mathcal{X}) - l(\mathbf{\Theta}^{(k-1)};\mathcal{X})$ is sufficiently small.

Appendix A Denseness

In this Appendix, we formulate the theorem stating that the class of mixtures of Erlang distributions with a common scale parameter is dense in the space of distributions on $\mathbb{R}^+$ (see Tijms (1994, p. 163)).

Theorem A.1. *The class of mixtures of Erlang distributions with a common scale parameter is dense in the space of distributions on* $\mathbb{R}^+$. *More specifically, let* $F(x)$ *be the cumulative* *distribution function of a positive random variable. Define the following cumulative distribution function of a mixture of Erlang distributions with a common scale parameter* $\theta > 0$*,*

$$F(x;\theta) = \sum_{j=1}^{\infty} \alpha_j(\theta) F(x; j, \theta) \, ,$$

where $F(x; j, \theta)$ *denotes the cumulative distribution function of an Erlang distribution with shape* j *and scale* θ*,*

$$F(x; j, \theta) = 1 - \sum_{n=0}^{j-1} e^{-x/\theta} \frac{(x/\theta)^n}{n!} \, ,$$

and the mixing weights are given by

$$\alpha_j(\theta) = F(j\theta) - F((j-1)\theta) \qquad \textit{for } j = 1, 2, \dots \, .$$

Then

$$\lim_{\theta \to 0} F(x;\theta) = F(x) \, ,$$

for each point x *at which* $F(\cdot)$ *is continuous.*

References

Lee, G. and Scott, C. (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816 – 2829.

Lee, S. C. and Lin, X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14(1):107.

McLachlan, G. and Peel, D. (2001). *Finite mixture models.* Wiley.

Tijms, H. C. (1994). *Stochastic models: an algorithmic approach.* Wiley.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be