The EM algorithm for loss distributions modelling

Sandra Teodorescu* **and Raluca Vernic**

Abstract. In this paper we present different approaches to model the loss distributions of two data sets from automobile liability insurance, with accent on the Estimation Maximization (EM) algorithm. A comparison study between those approaches is conducted.

Keywords: loss distributions, estimation maximization (EM) algorithm, automobile liability insurance.

1. Introduction

As specified in Klugman et al. (1998), „*in the most general sense, all of actuarial science is about loss distributions because that is precisely what an insurance agreement is all about*". The policy holder is paid a random amount (the loss) at a random future time. Hence, a loss distribution is considered to be the probability distribution of either the loss, or the amount paid from a loss event. Evaluating the loss distribution for an homogeneous portfolio is of great importance for the insurance company, because this distribution is involved in developing probability distributions for the aggregate loss, and therefore in evaluating ruin probabilities, reserves, benefits etc., or in establishing the influence of different deductibles.

In this paper we will consider different approaches to model the loss distributions of two data sets from automobile liability insurance. The data were kindly provided by two romanian insurance companies and consists of all the liability claims settled during year 2004 for an entire portfolio.

We will first present the Estimation Maximization (EM) algorithm used to evaluate maximum likelihood estimators for mixtures of normal densities (section 2). Then we will apply this algorithm for our data sets, conducting also a comparative study with other methods. In section 3 we study the first data set, while section 4 is dedicated to the second data set.

2. The EM algorithm

The EM algorithm is a popular tool for simplifying difficult maximum likelihood problems, see e.g. Dempster et al. (1977). This algorithm is designed for mixtures of normal distributions, and therefore it can be used when we notice that the distribution graph (e.g. histogram) presents a multi-modality. Then the number of modes should give an idea on the number of mixed distributions. In the following we will describe this algorithm for two and three mixtures of normal distributions.

2.1 EM algorithm for a two components mixture

Let us consider a sample $(y_1,...,y_N)$ from the random variable *D*. Assuming that the corresponding histogram reflects a bi-modality, then we can model *D* as a mixture of two normal random variables,

$$D = IC_1 + (1-I)C_2,$$

where

$$\begin{aligned} C_1 &\sim N(\mu_1, \sigma_1^2) \\ C_2 &\sim N(\mu_2, \sigma_2^2) \end{aligned},$$

* Ecological University of Bucharest; e-mail: sandra.teodorescu@ueb.ro

and $I \in \{0,1\}$, with $P(I=1)=r$. This representation is explicit: generate an $I \in \{0,1\}$ with probability *r* and then, depending on the outcome, deliver either C_1 or C_2.

Now let $\phi_i(x)$, $i=\overline{1,2}$, denote the normal density with parameters μ_i, σ_i^2. Then the density of *D* is

$$f_D(x) = r\phi_1(x) + (1-r)\phi_2(x).$$

In order to fit this model to the sample $(y_1,...,y_N)$, we must first estimate the parameters $r, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$. From Dempster et al. (1977), we have the following EM algorithm for a two-components Gaussian mixture:

Algorithm 1

1. Take initial guesses for the parameters, e.g. $\hat{r}=0{,}5$, $\hat{\mu}_1, \hat{\mu}_2$ taken at random from the observed data, and $\hat{\sigma}_1 = \hat{\sigma}_2$ equal to the overall sample variance.
2. *Expectation step*: compute the „responsabilities",

$$\hat{\gamma}_i = \frac{\hat{r}\phi_1(y_i)}{\hat{r}\phi_1(y_i) + (1-\hat{r})\phi_2(y_i)}, i = 1,...,N.$$

3. *Maximization step*: compute the weighted means and variances,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i y_i}{\sum_{i=1}^{N}\hat{\gamma}_i}, \quad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i (y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{N}\hat{\gamma}_i}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i) y_i}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)}, \quad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N}(1-\hat{\gamma}_i)(y_i - \hat{\mu}_2)^2}{\sum_{i=1}^{N}(1-\hat{\gamma}_i)},$$

and the mixing probability $\hat{r} = \frac{\sum_{i=1}^{N}\hat{\gamma}_i}{N}$.

4. Iterate steps 2 and 3 until convergence.

2.2 EM algorithm for a three components mixture

Based on the algorithm above, a generalization for three components mixtures is easy to obtain. Let us consider again a sample $(y_1,...,y_N)$ from the random variable *D*, but we will now assume that the corresponding histogram presents a tri-modality, hence we try to model *D* as a mixture of three normal random variables,

$$D = I_1 C_1 + I_2 C_2 + (1 - I_1 - I_2) C_3,$$

where

$$C_1 \sim N(\mu_1, \sigma_1^2)$$
$$C_2 \sim N(\mu_2, \sigma_2^2),$$
$$C_3 \sim N(\mu_3, \sigma_3^2)$$

and the bivariate distribution of the random variable (I_1, I_2) is given by
$P(I_1 = 1, I_2 = 0) = r_1$, $P(I_1 = 0, I_2 = 1) = r_2$, $P(I_1 = 0, I_2 = 0) = 1 - r_1 - r_2$, $0 < r_i < 1$, $i = 1,2$, and $r_1 + r_2 < 1$.

We will now denote by $\phi_i(x)$, $i = \overline{1,3}$, the normal density with parameters μ_i, σ_i^2, corresponding to the random variables C_i. Then the density of *D* writes

$$f_D(x) = r_1\phi_1(x) + r_2\phi_2(x) + (1 - r_1 - r_2)\phi_3(x).$$

We must estimate the parameters $r, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \mu_3, \sigma_3^2$. Based on Algorithm 1, the EM algorithm for a three-components Gaussian mixture becomes:

Algorithm 2

1. Take initial guesses for the parameters, $\hat{\mu}_i, \hat{\sigma}_i^2, i = 1,2,3$, $\hat{r}_i, i = 1,2$ (see suggestions in Algorithm 1).
2. *Expectation step*: compute the „responsabilities",

$$\hat{\gamma}_{i1} = \frac{\hat{r}_1\phi_1(y_i)}{\hat{r}_1\phi_1(y_i) + \hat{r}_2\phi_2(y_i) + (1 - \hat{r}_1 - \hat{r}_2)\phi_3(y_i)}, i = 1,..., N$$
$$\hat{\gamma}_{i2} = \frac{\hat{r}_2\phi_2(y_i)}{\hat{r}_1\phi_1(y_i) + \hat{r}_2\phi_2(y_i) + (1 - \hat{r}_1 - \hat{r}_2)\phi_3(y_i)}, i = 1,..., N$$

.

3. *Maximization step*: compute the weighted means and variances,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i1} y_i}{\sum_{i=1}^{N} \hat{\gamma}_{i1}}, \quad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i1}(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{N} \hat{\gamma}_{i1}}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i2} y_i}{\sum_{i=1}^{N} \hat{\gamma}_{i2}}, \quad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i2}(y_i - \hat{\mu}_2)^2}{\sum_{i=1}^{N} \hat{\gamma}_{i2}} \quad ,$$

$$\hat{\mu}_3 = \frac{\sum_{i=1}^{N} (1 - \hat{\gamma}_{i1} - \hat{\gamma}_{i2}) y_i}{\sum_{i=1}^{N} (1 - \hat{\gamma}_{i1} - \hat{\gamma}_{i2})}, \quad \hat{\sigma}_3^2 = \frac{\sum_{i=1}^{N} (1 - \hat{\gamma}_{i1} - \hat{\gamma}_{i2})(y_i - \hat{\mu}_3)^2}{\sum_{i=1}^{N} (1 - \hat{\gamma}_{i1} - \hat{\gamma}_{i2})}$$

and the mixing probabilities $\hat{r}_1 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i1}}{N}$, $\hat{r}_2 = \frac{\sum_{i=1}^{N} \hat{\gamma}_{i2}}{N}$.

4. Iterate steps 2 and 3 until convergence.

3. Modelling a loss distribution for the first data set

Unfortunately, this data set consists of only 69 settled claims during 2004. This is not enough for a thorough statistical analisys, but one can get a general ideea. The main empirical characteristics of this data set are

Expected value=17.837.681
Standard deviation= 18.242.961
Skewness=2,09
Standard Error Skewness=0,28
Kurtosis= 5,51
Standard Error Kurtosis= 0,57
Maximum value=100.000.000
Minimum value=1.100.000

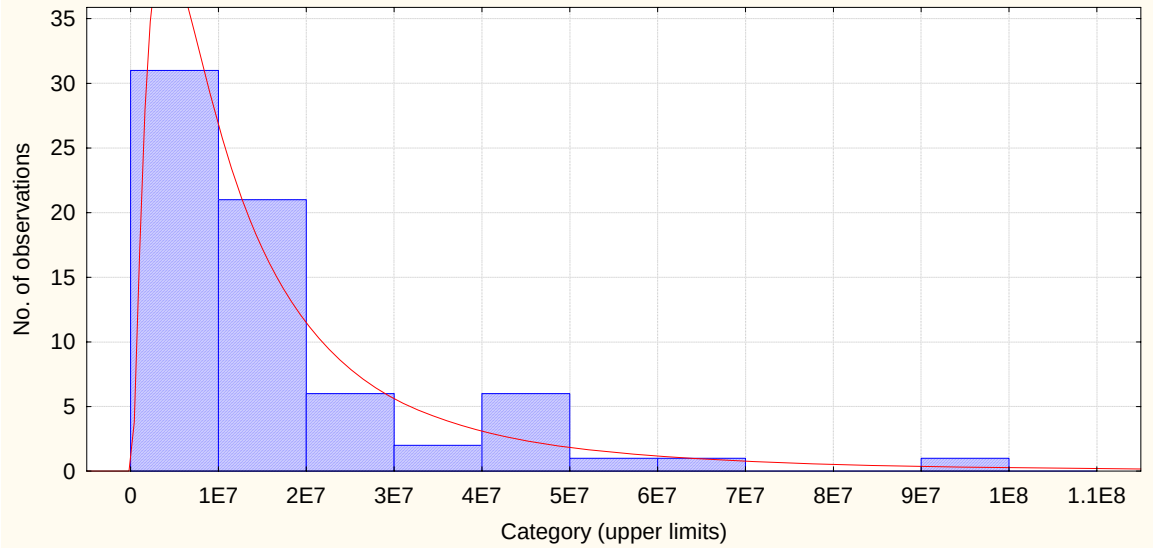
In the following, we will consider three different approaches to fit a distribution to this data set.

3.1 A first approach

Using Statistica software, we tried to fit several continuous distributions that are usually considered when modeling claims. For details on these distributions see e.g. Kotz et al. (2000) or Kaas et al. (2001). As a fit measure we used the *p*-value (or *p*-level) given by Statistica, with the following meaning: it represents the probability of error that is involved in accepting our observed result as valid. In many areas of research, the upper-limit *p*-value accepted is 0,05, that is a *p*-value $p \le 0,05$ is considered at trust limit. A significant *p*-value should be $p \le 0,01$, while $p \le 0,005$ or $p \le 0,001$ are highly significant.

Back to our data, we first tried to fit a Log-Normal distribution, and the result is given in Figure 1.

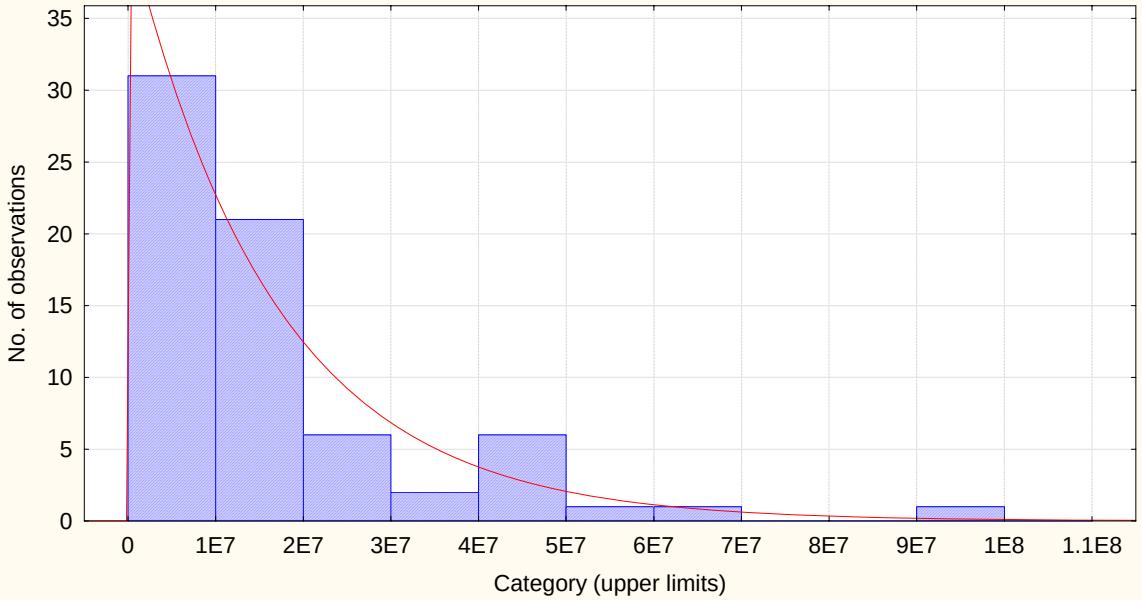
Figure 1. Data set I
Log-normal distribution in red line



With a *p*-value *p*=0,28 , it is clear that the Log-Normal distribution must be rejected.

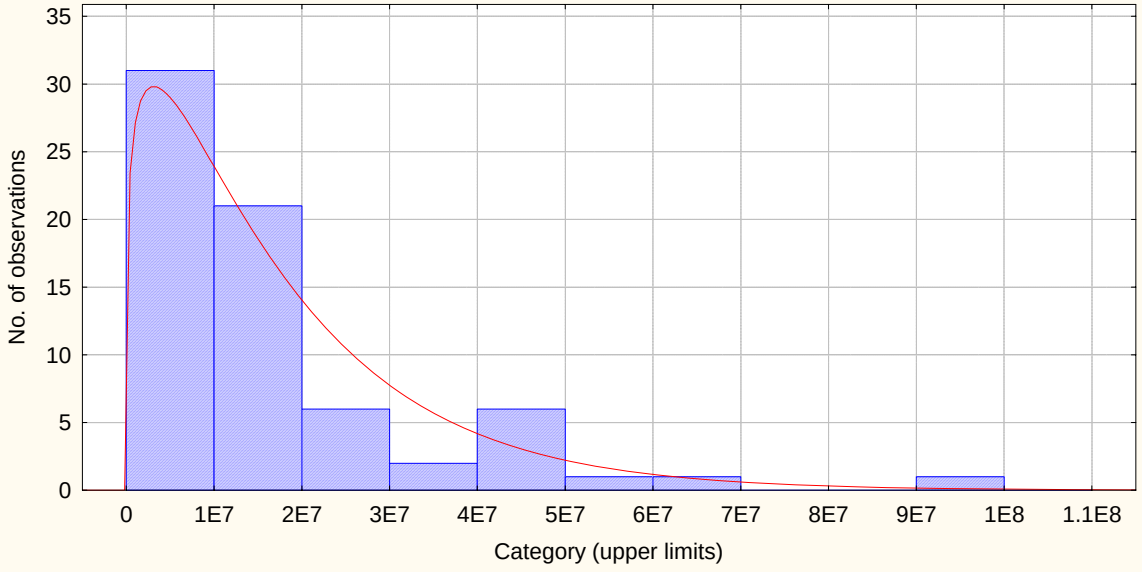
Secondly, we tried to fit an Exponential distribution, see Figure 2. Again, the *p*-value *p*=0,15 is too high, so we reject this distribution also.

Figure 2. Data set I
Exponential distribution in red line



Finally, we tried to fit a Gamma distribution, as in Figure 3. This time, the *p*-value p=0,044 is at the trust limit, so this distribution can be considere as a valid candidate for this data set if we cannot find a better one.

Figure 3. Data set I
Gamma distribution in red line



3.2 A second approach

Looking closer at the data histogram, one can notice that the claims seem to be of two kinds: lower costs and higher costs. This is why we decided to try to fit a mixture of two normal distributions, one normal distribution for the lower costs considered independent, identically distributed (i.i.d.) like the random variable C_1, and a second normal distribution for the higher

costs, also i.i.d. like C_2. Then denoting by *D* the claim random variable corresponding to all the data, it will be of the form

$$D = IC_1 + (1-I)C_2,$$

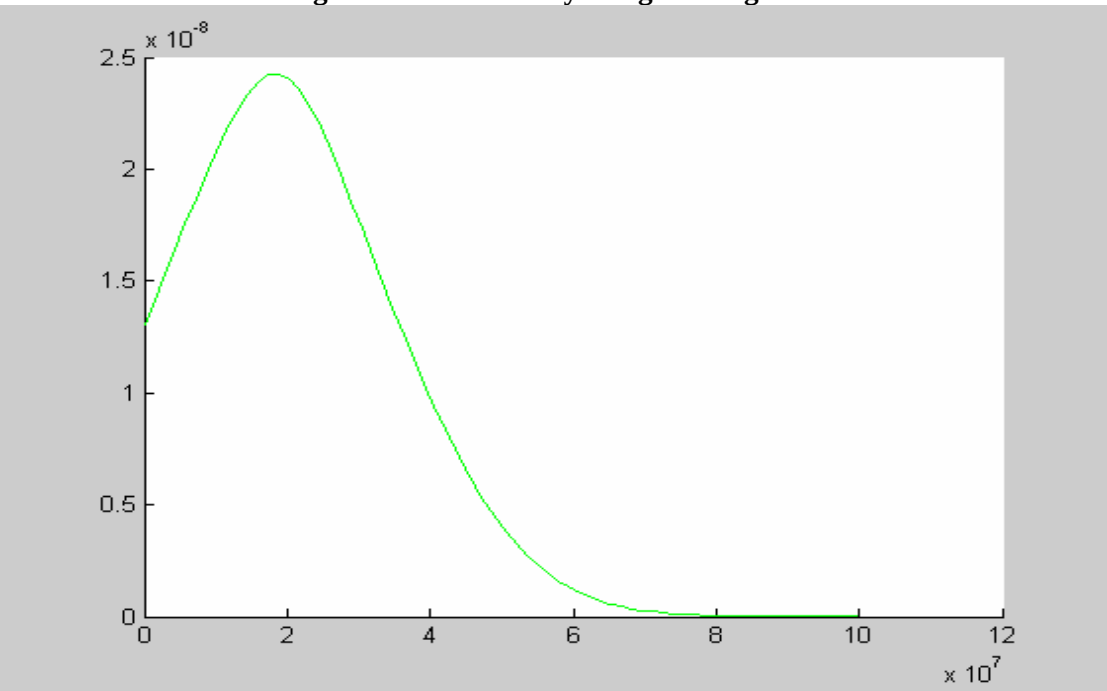
where *I* is a *Bernoulli* random variable, i.e. $I:\begin{pmatrix} 0 & 1 \\ 1-r & r \end{pmatrix}$. We recognize here the model from section 2.1, with the interpretation: if $I=1$, then the claim *D* equals C_1, and if $I=0$, then the claim *D* will be equal to C_2. Therefore, using the notation in section 2.1, we can apply Algorithm 1 to try to fit to the data a density of the form

$$f_D(x) = r\phi_1(x) + (1-r)\phi_2(x).$$

The fitted density obtained using MATLAB can be seen in Figure 4, while the parameters are

$$r = 0,95539$$
$$\mu_1 = 1,7782 \bullet 10^7 \quad \sigma_1^2 = 3,1029 \bullet 10^{14}$$
$$\mu_2 = 1,9038 \bullet 10^7 \quad \sigma_2^2 = 4,5285 \bullet 10^{13}$$

Figure 4. Fitted density using EM Algorithm 1



3.3 A third approach

Using the same notation as in section 3.2, we assume again that the distribution of *D* is a mixture of two distributions, just that this time those two distributions will not be considered of normal type. Looking at the histogram, a realistic assumption would be that the lower costs are exponentially distributed, while for the higher costs we can choose a Pareto distribution, since Pareto is classically used for extreme value costs. The problem is to establish the proportion *r* of the lower costs from the total costs. Empirically, we choose the first 59 values as the lower costs and the last 10 values as the higher costs, so we will have *r*=59/69=0,855.

Hence, we have $C_1 \sim Exponential(\theta)$ with $F_{C_1}(x) = 1 - e^{-\theta x}$ and a maximum likelihood estimated value for the parameter $\theta_{VM} = 8,641 \cdot 10^{-8}$.

Also, taking $C_2 \sim Pareto(a, \alpha)$ with $F_{C_2}(x) = \begin{cases} 1 - \left(\frac{a}{x}\right)^{\alpha}, & x > a \\ 0, & x \le a \end{cases}$, we choose $a = 40.000.000$, while the maximum likelihood estimation for the other parameter is $\alpha = 3,5852$.

A better fitting of this Exponential-Pareto mixture could be obtained using an adapted EM algorithm with starting values for the parameters given by the ones above. This will be subject for further research.

In conclusion, we expect that this Exponential-Pareto mixture is better that the Normal-Normal one presented in section 3.2, and better that the Gamma distribution fitted in section 3.1.

4. Modelling a loss distribution for the second data set

This second data set consists of 1161 settled claims during 2004, which is better than for the first data set. The main empirical characteristics of this data set are

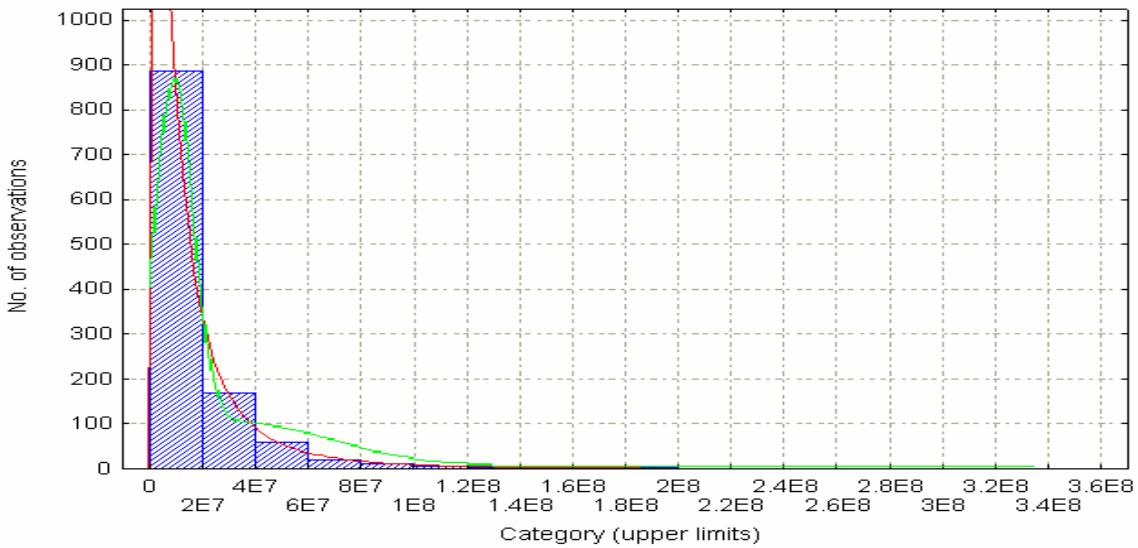
Expected value=17.126.337,4
Standard deviation=24.267.282,3
Skewness=4,62
Standard Error Skewness=0,07
Kurtosis= 32,80
Standard Error Kurtosis= 0,14
Maximum value=310.000.000
Minimum value=9.000

In the following, we will consider two different approaches to fit a distribution.

4.1 The first approach

As before, using Statistica software, we tried to fit several continuous distributions and we noticed that the „best" fit is given by the Log-normal distribution (see Figure 5, red line), but since $p = 0,77880 > 0,05$, we have to reject this distribution and try another model.

Figure 5. Data set II
Log-normal distribution in red line EM density in green line



4.2 The second approach

Looking closer at the data histogram, we noticed that this time the claims seem to be of three kinds: lower costs, medium costs and higher costs. Hence, we try to fit a mixture of three normal distributions, one normal distribution for each kind of costs: lower costs of random variable C_1, medium costs of random variable C_2, and higher costs of random variable C_3. Then denoting by *D* the claim random variable corresponding to all the data, it will be of the form

$$D = I_1 C_1 + I_2 C_2 + (1 - I_1 - I_2) C_3,$$

so that we recognize here the model from section 2.2, with the interpretation: if $I_1 = 1$, then the claim *D* equals C_1, if $I_2 = 1$ then *D* equals C_2, and if $I_1 = I_2 = 0$, then *D* will be equal to C_3. Therefore, using the notation in section 2.2, we can apply Algorithm 2 to try to fit to the data a density of the form

$$f_D(x) = r_1 \phi_1(x) + r_2 \phi_2(x) + (1 - r_1 - r_2) \phi_3(x).$$

The fitted density obtained using MATLAB can be seen in Figure 5 (green line), while the estimated parameters are:

$$r_1 = 0,38576 \qquad r_2 = 0,22043$$
$$\mu_1 = 9,1247 \bullet 10^6 \quad \sigma_1^2 = 4,6276 \bullet 10^{13}$$
$$\mu_2 = 8,1295 \bullet 10^6 \quad \sigma_2^2 = 4,0952 \bullet 10^{13}$$
$$\mu_3 = 3,0001 \bullet 10^7 \quad \sigma_2^2 = 1,242 \bullet 10^{15}$$

In conclusion, we expect that this Normal mixture is better that the Log-normal distribution presented in section 4.1, and we also want to improve it by considering more than three normal distributions in the mixture. This will be subject for further research also.

References

1. Dempster, A.P.; Laird, M.; Rubin, D.B. (1977) - Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B 39 (1) , 1-38

2. Kaas, R.; Goovaerts, M.; Dhaene, J.; Denuit, M. (2001) – Modern Actuarial Risk Theory. *Kluwer, Boston.*

3. Klugman, S.A.; Panjer, H.H.; Willmot, G. (1998) – Loss Models. From data to decisions. *Wiley-Interscience, New York.*

4. Kotz, S.; Balakrishnan, N.; Johnson, N. L. (2000) - Continuous Multivariate Distributions. Vol. 1. Models and applications. *Wiley-Interscience, New York, 2nd edition.*