EM Algorithm

Cornelia Gruber

10th January 2021

1 Introduction & Data Description

The purpose of this paper is to determine the most plausible parameters of a distribution and their confidence intervals given the observed data. We have 300 observations of a one dimensional, non-negative variable. For the provided data the minimal observed value is 0.38, the first quartile is 1.45, the median is 2.16, the mean value is 2.37, the third quartile is 3.10 and the maximal observed value is 6.82. The density function of the observed data can be found in Figure 1 (black line) and as a reference the density function where the parameters were estimated by the EM algorithm is drawn in red ($\delta = 2.37, p = 1$).

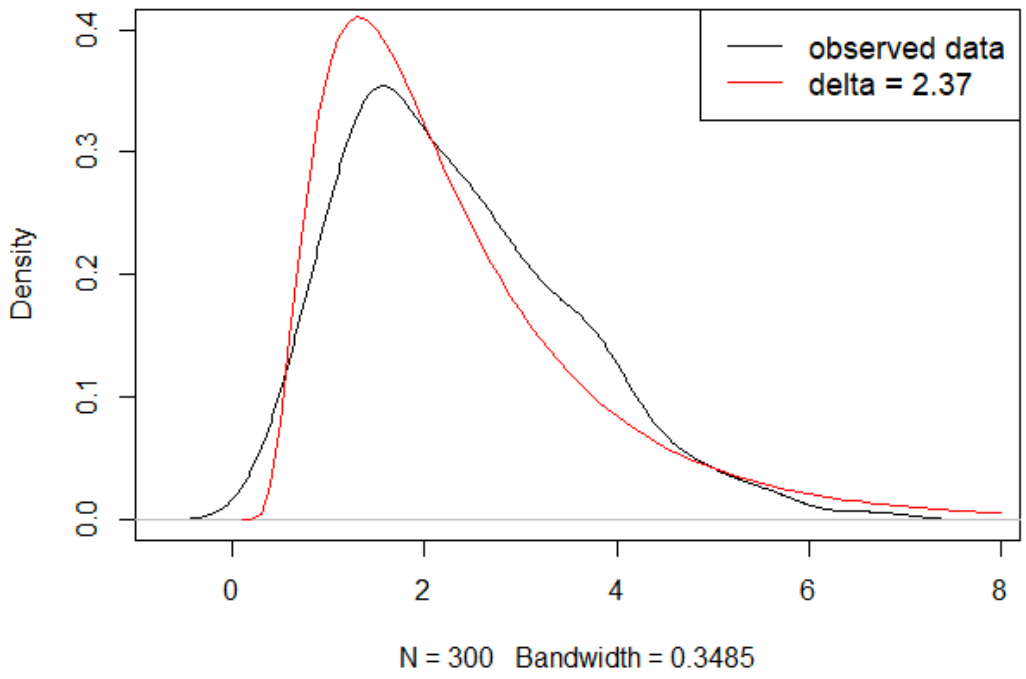

Figure 1: Density of observed data in black, density with estimated parameters in red ($\delta = 2.37, p = 1$)

2 Method Description

A common method to find the parameters (δ) of a distribution that best fits the observed data is maximum likelihood. For a distribution, where the data consist of n observations, $x = (x_1, \ldots, x_n)$, the likelihood function is defined as

$$L(\delta) = \prod_{i=1}^{n} f(x_i|\delta) \tag{1}$$

where f is the density of the particular distribution. However, oftentimes to simplify calculations the log-likelihood

$$l(\delta) = log(L(\delta)) = \sum_{i=1}^{n} logf(x_i|\delta) \tag{2}$$

is used instead.

In our example we have a mixture model, i.e. a distribution with k components, where we want to find the best parameters and need to maximize the likelihood. This is however not possible in an analytical way and we make use of an Estimation-Maximization (EM) algorithm. In our case the distribution function for sub-population j is given as

$$f(x|\delta_j) = \frac{\delta_j}{\sqrt{2\pi}} exp(\delta_j) x^{-3/2} exp\left(-\frac{1}{2}\left(\frac{\delta_j}{x} + x\right)\right), \quad x, \delta_j > 0 \tag{3}$$

and the full distribution is

$$f(x) = \sum_{j=1}^{k} p_j f(x|\delta_j). \tag{4}$$

The log likelihood is thus

$$l(\delta) = \sum_{i=1}^{n} log\left(\sum_{j=1}^{k} p_j f(x_i|\delta_j)\right) \tag{5}$$

2.1 EM-Algorithm

The EM algorithm is especially useful in case of missing data, e.g when the membership of each observation x_i to the sub-populations $1, \ldots, k$ is not known. We need to find the mixing probabilities $p_1, \ldots, p_k$ as well as the parameters $\delta_1, \ldots, \delta_k$. If the class membership for each observation is known, the parameters can be easily found. That is why we use the random variable Z_{ij} which indicates whether observation i belongs to class/sub-population j ($Z_{ij} = 1 \iff i$ belongs to class j). The complete data likelihood $l(\delta|x, Z)$ is then given by

$$\sum_{i=1}^{n}\sum_{j=1}^{k} z_{ij} log(p_j) + z_{ij}\left(log(\delta_j) - log(\sqrt{2\pi}) + \delta_j - \frac{3}{2} log(x_i) - \frac{1}{2}\left(\frac{\delta_j}{x_i} + x_i\right)\right) \tag{6}$$

In general during the EM-Algorithm two steps, the estimation and maximization step (E and M step), are repeated until convergence or lack of improvement. The "E-step" consists of the estimation of the expected value $w_{ij} = E(Z_{ij}|x_i, p, \delta)$. To be precise, we estimate w_{ij} by

$$\textbf{E-step:} \qquad w_{ij} = \frac{p_j f(x_i|\delta_j)}{\sum_{j=1}^{k} p_j f(x_i|\delta_j)} \tag{7}$$

Now that we have intermediate values for w_{ij} we can maximize the likelihood, which is therefor called the "M-Step". In detail we update the values for p and δ:

$$\textbf{M-step:} \qquad p_j^{new} = \frac{\sum_{i=1}^{n} w_{ij}}{n} \quad and \quad \delta_j^{new} = \frac{\sum_{i=1}^{n} w_{ij} x_i}{\sum_{i=1}^{n} w_{ij}} \tag{8}$$

Both the E and the M step are repeated until a convergence criterion is met. Commonly used criteria are either based on the change in parameters or the change in likelihood, the latter is used in our example. More precisely we stop iterating the steps when:

$$\left| \frac{l^{t+1} - l^t}{l^{t+1}} \right| \leq 0.000000001 \tag{9}$$

where l^t describes the likelihood at step t.

2.2 Metropolis-Hastings Sampling

In order to draw samples from a distribution f with Metropolis-Hastings we need a proposal density $q(x, y)$ from which we can easily sample from. The steps are then

1. generate candidate value $y \sim q(x, .)$
2. calculate $\alpha(x, y) = min\{\frac{f(y)q(y,x)}{f(x)q(x,y)}, 1\}$
3. accept new value y with probability α, keep old value if y is rejected

Those steps are repeated until the desired amount of samples is reached. In our example we use a Normal-distribution with a standard deviation of 4 in order to have a mean acceptance rate of around 30%. Now to get independent and identically distributed samples we need to ensure the chain of generated values (Markov-chain) reached its stationary distribution (our target distribution f). Therefor we discard the first 1000 draws (burn-in period) and then take every 20th value out of the Markov-chain resulting from the Metropolis-Hastings Sampling.

2.3 Bootstrapping

The bootstrapping procedure that was used in this paper proceeds as follows:

1. calculate EM-estimate for δ and p from initial sample

2. generate B bootstrap samples following the distribution from Equation (4) with δ and p from step 1 by using Metropolis-Hastings

3. calculate EM-estimate for δ and p from bootstrapped samples

2.4 Percentile confidence interval

The percentile confidence interval is given by

$$CI = [\kappa_{\alpha/2}, \kappa_{1-(\alpha/2)}] \tag{10}$$

where κ_α denotes the empirical α quantile of the bootstrapped values. In our case the bootstrapped values are the parameters δ and p of the distribution.

3 Results

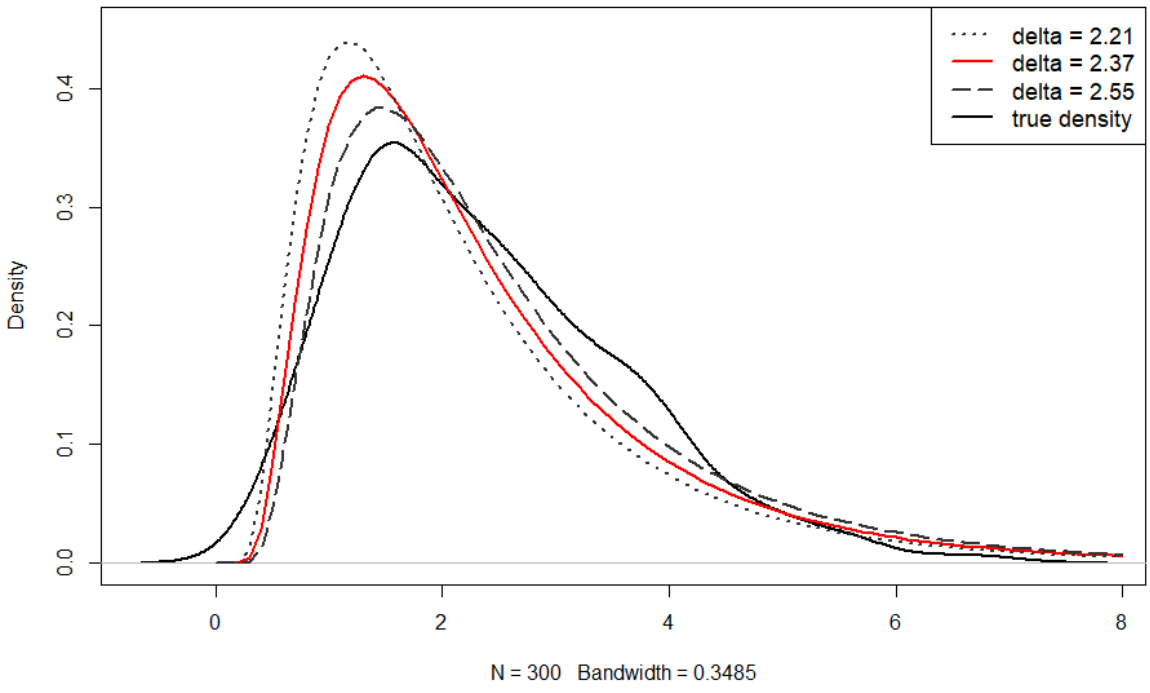

Figure 2: Density of observed data in black, EM-estimated density with $\delta = 2.37$ in red, density for lower CI-value $\delta = 2.21$ dotted and density for upper CI-value $\delta = 2.55$ dashed

Now if we apply this EM-algorithm to the provided data the most likely setting was a single group (k = 1) where obviously 100% of the data points belong to ($p = 1$) and a δ of 2.37. The estimated density can be found in red in Figure 1. As a rule of thumb, start with a large k and count how many unique values are estimated, and take that as

the most likely number of classes. Estimation with two, three, or even five groups (k = 2,3,5) leads to a random assignment of the classes, where each class has a δ of 2.37. Therefor the true data generating process probably only had one class and we continue with $k = 1, p = 1, \delta = 2.37$.

To asses the variability of the result a parametric bootstrap was performed, where B = 1000 bootstrap samples were drawn according to a Metropolis Hastings sampling procedure from the distribution given in Equation (3) with $\delta = 2.37$. When now calculating the 95%-percentile confidence intervals of the bootstrapped parameters we get:

$$\text{CI for } \delta = [2.21, 2.55] \tag{11}$$

The densities when using the lower and the upper end of the confidence interval can be found in Figure 2.