ISSN: 1935-7524

Estimation of completely monotone and k-monotone distribution functions

Piet Groeneboom

Delft University of Technology, Building 28, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands.
e-mail: P.Groeneboom@tudelft.nl

Abstract: We study maximum likelihood estimation of k-monotone and completely monotone distribution functions.

AMS 2000 subject classifications: Primary 62G09, 62N01.
Keywords and phrases: k-monotone distributions, completely monotone distributions, maximum likelihood estimation.

1. EM algorithm for the MLE for completely monotone distributions

Least squares estimation of a completely monotone probability mass function is discussed in [1], where further references to the literature can be found. Here we consider maximum likelihood estimation of the probability mass function. The log likelihood function is given by:

$$\sum_{i=1}^{n} \log \int_0^1 (1-\alpha)\alpha^{x_i}\, dF(\alpha),$$

where F is the mixing distribution. In particular, if we restrict the estimation to probability mass functions on a grid $\{\alpha_1, \ldots, \alpha_m\}$, where $0 < \alpha_1 < \cdots < \alpha_m < 1$, we have to maximize

$$\sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} (1-\alpha_j)\alpha_j^{x_i} q_j \right)$$

over point masses q_j such that $\sum_{j=1}^{m} q_j = 1$. Note that the function is concave in $(q_1, \ldots, q_m)$, but not concave in $(\alpha_1, \ldots, \alpha_m)$. But in fact we would have to maximize over $(q_1, \ldots, q_k)$, $(\alpha_1, \ldots, \alpha_m)$ and m.

Working on a fixed grid $\{\alpha_1, \ldots, \alpha_m\}$, and introducing a Lagrange multiplier, we have to minimize

$$\phi_\lambda(q_1, \ldots, q_m) = -n^{-1} \sum_{i=1}^{n} \log \left(\sum_{j=1}^{m} (1-\alpha_j)\alpha_j^{x_i} q_j \right) + \lambda \sum_{j=1}^{m} q_j$$

Taking the derivative w.r.t. q_j, we get

$$\frac{\partial}{\partial q_j} \phi_\lambda(q_1, \ldots, q_m) = -n^{-1} \sum_{i=1}^{n} \frac{(1-\alpha_j)\alpha_j^{x_i}}{\sum_{k=1}^{m} (1-\alpha_k)\alpha_k^{x_i} q_k} + \lambda,$$

implying $\lambda = 1$. So the self-consistency equation for the q_j's is:

$$q_j = n^{-1} \sum_{i=1}^{n} \frac{(1-\alpha_j)\alpha_j^{x_i} q_j}{\sum_{k=1}^{m} (1-\alpha_k)\alpha_k^{x_i} q_k}\,.$$

This induces the EM iterations

$$q_j^{new} = n^{-1} \sum_{i=1}^{n} \frac{(1-\alpha_j)\alpha_j^{x_i} q_j^{old}}{\sum_{k=1}^{m} (1-\alpha_k)\alpha_k^{x_i} q_k^{old}}\,.$$

The minimization is done on the cone $\mathbb{R}^m_+$, so the Fenchel duality conditions are:

$$\frac{\partial}{\partial q_j}\phi_\lambda(q_1,\ldots,q_m) \geq 0, \qquad j=1,\ldots,m, \tag{1.1}$$

and

$$\sum_{j=1}^m q_j \frac{\partial}{\partial q_j}\phi_\lambda(q_1,\ldots,q_m) = 0. \tag{1.2}$$

We iterate the EM algorithm until these conditions are satisfied at a certain accuracy (like 10^{-10}).

2. Growing support algorithm

The growing support algorithm works by starting with a minimal set of support points to make the log likelihood finite and then letting the support grow until adding further support points does not seem to make the likelihood larger. In the present case one could start with one point, say $\alpha \in (0,1)$ which has mass 1. Then the criterion function ϕ_λ of the preceding section is given by

$$\phi_1(\alpha;q) = -n^{-1}\sum_{i=1}^n \log\{(1-\alpha)\alpha^{x_i}\} + q, \tag{2.1}$$

where we give the criterion function two sets of arguments: the locations of the masses $\alpha_1,\ldots\alpha_m$ and the values of the masses at these points: $q_1,\ldots,q_m$. In this case we only have $\alpha_1=\alpha$ and $q_1=1$. We can now optimize for one support point by computing the α minimizing (2.1). As an example, if $n=2$ we get as solution

$$\alpha = \frac{x_1+x_2}{2+x_1+x_2}.$$

If not all x_i's are equal to 0, we can expect solutions for α in the open interval $(0,1)$.

Generally, we have the criterion function

$$\phi_m(\alpha_1,\ldots,\alpha_m;q_1,\ldots,q_m) = -n^{-1}\sum_{i=1}^n \log\left(\sum_{j=1}^m (1-\alpha_j)\alpha_j^{x_i}q_j\right) + \sum_{j=1}^m q_j,$$

where we assume $q_i \in (0,1)$, $i=1,\ldots,m$. We assume that $0<\alpha_1<\cdots<\alpha_m<1$. A necessary criterion for having a minimum for m variables is that

$$\frac{\partial}{\partial \alpha_j}\phi_1(\alpha_1,\ldots,\alpha_m;q_1,\ldots,q_m) = 0, \qquad j=1,\ldots,m, \tag{2.2}$$

and

$$\frac{\partial}{\partial q_j}\phi_1(\alpha_1,\ldots,\alpha_m;q_1,\ldots,q_m) = 0, \qquad j=1,\ldots,m, \tag{2.3}$$

The growing support algorithm now tests whether we can improve by adding a further point α_{m+1}, considering the criterion function

$$\phi_{m+1}(\alpha_1,\ldots,\alpha_{m+1};q_1,\ldots,p_{m+1}) = -n^{-1}\sum_{i=1}^n \log\left(\sum_{j=1}^{m+1} (1-\alpha_j)\alpha_j^{x_i}q_j\right) + \sum_{j=1}^{m+1} q_j,$$

That is, we look for vectors $(\alpha_1,\ldots,\alpha_{m+1})$ and $(q_1,\ldots,p_{m+1})$, both in $(0,1)^{m+1}$, where (2.2) and (2.3) are satisfied, with m replaced by $m+1$, and check whether there is an sufficient improvement of ϕ_{m+1} w.r.t. ϕ_m according to some pre-determined criterion.

3. Smooth functionals

We relate the moments of F to the probabilities of the mixing distribution and let $m_k = \int_0^1 \alpha^k \, dF(\alpha)$, $k = 0, 1, \ldots$ denote the moments of F. Let X have the discrete distribution, specified by

$$q_k = \mathbb{P}\big\{X = k\big\} = \int_0^1 (1-\alpha)\alpha^k \, dF(\alpha), \qquad k = 0, 1, \ldots .$$

Then

$$q_k = \int_0^1 \alpha^k \, dF(\alpha) - \int_0^1 \alpha^{k+1} \, dF(\alpha) = m_k - m_{k+1}, \qquad k = 0, 1, \ldots$$

Hence, assuming $m_k \to 0$, we get:

$$\mathbb{P}\{X \geq k\} = m_k.$$

For the uniform distribution F we get:

$$\mathbb{P}\{X \geq k\} = \int_0^1 \alpha^k \, d\alpha = \frac{1}{k+1}\,, \qquad q_k = \frac{1}{k+1} - \frac{1}{k+2} = \frac{1}{(k+1)(k+2)}\,, \qquad k = 0, 1, \ldots$$

Let A have the distribution function F. We have:

$$\left[L_F a\right](k) = \mathbb{E}\{a(A)|X = k\} = \frac{\int_0^1 a(\alpha)(1-\alpha)\alpha^k \, dF(\alpha)}{\int_0^1 (1-\alpha)\alpha^k \, dF(\alpha)}, \qquad k = 0, 1, \ldots$$

The adjoint of L_F is given by:

$$\left[L_F^* b\right](\alpha) = \mathbb{E}\{b(X)|A = \alpha\} = \sum_{k=0}^{\infty} (1-\alpha)\alpha^k b(k), \qquad \alpha \in \text{support}(dF)^\circ.$$

where $\text{support}(dF)^\circ$ denotes the interior of the support of dF. For estimating the ith moment of F, we get the equation

$$\left[L_F^* b\right](\alpha) = \sum_{k=0}^{\infty} (1-\alpha)\alpha^k b(k) = \alpha^i - m_i, \qquad \alpha \in \text{support}(dF)^\circ.$$

The efficient influence function a_F is now given by solving in a the equations:

$$\left[L_F a\right](k) = \frac{\int_0^1 a(\alpha)(1-\alpha)\alpha^k \, dF(\alpha)}{\int_0^1 (1-\alpha)\alpha^k \, dF(\alpha)} = b(k), \qquad k = 0, 1, \ldots$$

So we get in fact the equations:

$$\sum_{k=0}^{\infty} (1-\alpha)\alpha^k \frac{\int_0^1 a(\beta)(1-\beta)\beta^k \, dF(\beta)}{\int_0^1 (1-\beta)\beta^k \, dF(\beta)} = \alpha^i - m_i, \qquad \alpha \in \text{support}(dF)^\circ.$$

This also gives the (efficient) asymptotic variance of the moment functionals for the (plugged in) MLE. See for similar computations Section 10.2 of [2], but note that proofs still have to be given for the present situation.

3.1. Example

As an example, we consider the following distribution for F: F corresponds to a discrete distribution on the set $\{\alpha_1, \ldots, \alpha_4\}$ with probability masses q_i on α_i, given by:

$$\alpha_1 = 0.15, \quad \alpha_2 = 0.25, \quad \alpha_3 = 0.4, \quad \alpha_4 = 0.75, \qquad q_1 = 0.3, \quad q_2 = 0.15, \quad q_3 = 0.25, \quad q_4 = 0.3, \tag{3.1}$$

where (as before) we use the notation $q_i = \mathbb{P}\{U = \alpha_i\}$, where U has the mixing distribution F.
We have to solve:

$$\sum_{k=0}^{\infty}(1-\alpha)\alpha^k \frac{\int_0^1 a(\beta)(1-\beta)\beta^k\, dF(\beta)}{\int_0^1 (1-\beta)\beta^k\, dF(\beta)} = \alpha - m_1 \qquad \alpha \in \text{support}(dF_0)^\circ,$$

where the first moment $m_1 = 0.4075$ in the present case. This yields 4 linear equations in 4 unknowns for $a(\alpha_i)$, $i = 1, \ldots, 4$. In this case we get:

$$\int_0^1 a(\alpha)(1-\alpha)\alpha^k\, dF(\alpha) = \sum_{i=1}^{4} a(\alpha_i)(1-\alpha_i)\alpha_i^k q_i.$$

Let $a_i = a(\alpha_i)$. Then we can write the equations in the form:

$$\sum_{i=1}^{4} a_i \sum_{k=0}^{\infty} \frac{(1-\alpha_j)\alpha_j^k(1-\alpha_i)\alpha_i^k q_i}{\sum_{l=1}^4 (1-\alpha_l)\alpha_l^k q_l} = \alpha_j - m_1, \qquad j = 1, \ldots, 4.$$

This has as solution

$$a_1 = -7.33729, \qquad a_2 = 16.93362, \qquad a_3 = -2.16622, \qquad a_4 = 0.67567. \tag{3.2}$$

Following the treatment on p. 288 of [2] and using the notation there, we get:

$$\theta_F(k) = \frac{\int_0^1 a(\alpha)(1-\alpha)\alpha^k\, dF(\alpha)}{\int_0^1 (1-\alpha)\alpha^k\, dF(\alpha)} = \frac{\sum_{i=1}^4 a_i(1-\alpha_i)\alpha_i^k q_i}{\sum_{i=1}^4 (1-\alpha_i)\alpha_i^k q_i} = \frac{\sum_{i=1}^4 a_i(1-\alpha_i)\alpha_i^k q_i}{p_k},$$

where $p_k = \sum_{i=1}^4 (1-\alpha_i)\alpha_i^k q_i$ is the induced probability in the observation space that the observation is equal to k, and where the efficient asymptotic variance is given by:

$$\|\theta_F\|^2_{P_F} = \sum_{k=0}^{\infty} \frac{\left\{\sum_{i=1}^4 a_i(1-\alpha_i)\alpha_i^k q_i\right\}^2}{p_k},$$

where P_F is the induced (by F) measure in the observation space. Evaluating this for 100 terms in the series of terms with index k and using (3.2), we get:

$$\|\theta_F\|^2_{P_F} \approx 0.240236.$$

So we get the following result

Proposition 3.1. *Let $\hat{F}_n$ be the MLE of F in the model, specified by (3.1). Then we have:*

$$\sqrt{n}\left\{\int_0^1 \alpha\, d\hat{F}_n(\alpha) - \int_0^1 \alpha\, dF(\alpha)\right\} \xrightarrow{\mathcal{D}} N\left(0, \|\theta_F\|^2_{P_F}\right), \tag{3.3}$$

where

$$\|\theta_F\|^2_{P_F} \approx 0.240236. \tag{3.4}$$

The variance $\|\theta_F\|^2_{P_F}$ is the information lower bound for the asymptotic variance in this estimation problem.

Remark 3.1. This line of argument can be extended to any moment of the mixing distribution.

We tested the theory by generating 1000 samples from the mixing distribution of the sample and generating geometric random variables with parameter α if α was chosen by the mixing distribution. The MLE was computed by the growing support algorithm of Section 2. The variances times n were given by 0.23665 for $n = 1000$ and 0.24017 for $n = 10,000$, where in particular the last value is very close to the value of the asymptotic lower bound in (3.4). The estimates of the mean were 0.40667 and 0.40758, for $n = 1000$ and $n = 10,000$, respectively. A boxplot of the simulations is shown in Figure 1.

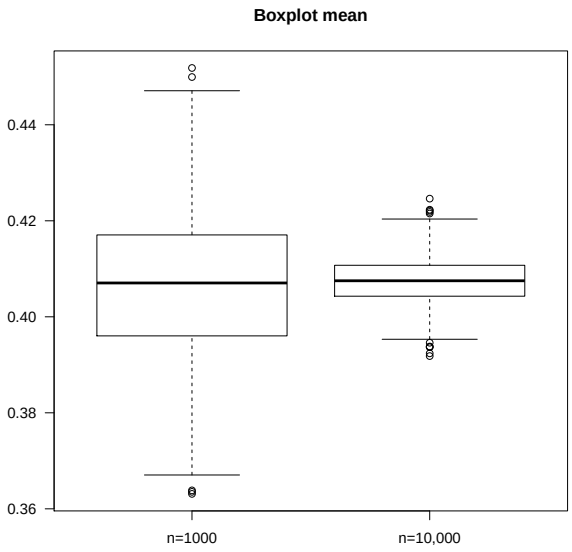

Fig 1: Boxplot of $\int_0^1 \alpha \, d\hat{F}_n(\alpha)$ for the model (3.1), where $\hat{F}_n$ is the MLE of the mixing distribution, for 1000 replications and sample sizes $n = 1000$ and $n = 10,000$.

4. The 1-monotone case

We can also consider the k-monotone distributions as mixtures. As an example, we consider the case $k = 1$. When $k = 1$, the discrete mixture density has the representation

$$p_F(x) = \int_{y\in[0,\infty)} \frac{1}{1+y} 1_{[0,y]}(x) \, dF(y),$$

and the log likelihood changes to:

$$\sum_{i=1}^n \log \left\{ \int_{y\in[0,\infty)} \frac{1}{1+y} 1_{[0,y]}(X_i) \; dF(y) \right\} = \sum_{i=1}^n \log \left\{ \int_{y\geq X_i} \frac{1}{1+y} \, dF(y) \right\},$$

where F is the distribution function of a probability measure on the integers $\mathbb{Z}_+$.

We consider here the estimation of F. It is clear that p_F is a decreasing density and all decreasing densities have a representation of this type. As an example we could take F to be the distribution function of the Poisson distribution with parameter μ:

$$q_j = P_F\{U = j\} = \frac{e^{-\mu}\mu^j}{j!}, \qquad j = 0, 1, \ldots \tag{4.1}$$

So in this case we get for the induced distribution in the observation space:

$$P_F\{X = i\} = \sum_{k=i}^{\infty} \frac{1}{1+k} \frac{e^{-\mu}\mu^k}{k!} = \frac{\Gamma(1+i) - \int_\mu^\infty x^i e^{-x}\, dx}{\mu\, \Gamma(1+i)}\,. \tag{4.2}$$

This means that an observation is of the form

$$X_i, \qquad i = 1, \ldots, n,$$

where, conditionally on Y_i, X_i is uniformly chosen on $\{0, \ldots, Y_i\}$, and Y_i is Poisson(μ). A plot of the probabilities (4.2), made in Mathematica, is given in Figure 2.

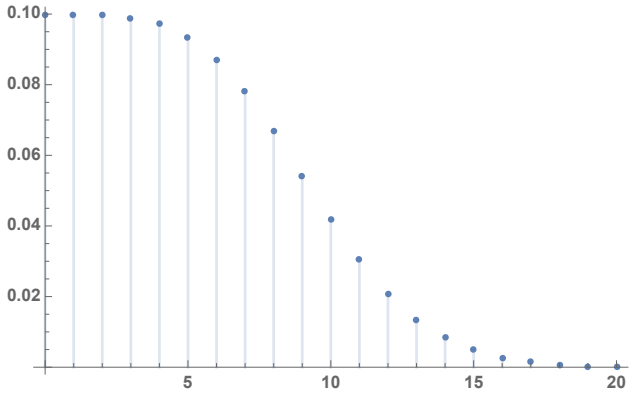

Fig 2: Plot of P_F in (4.2) for $\mu = 10$.

Generally the criterion function to be minimized is given by:

$$-\sum_{i=1}^{n} \log \left\{ \sum_{j \geq X_i} \frac{q_j}{1+j} \right\} + \sum_{j \geq 0} q_j$$

where q_j is in fact given by (4.1). As before we work with a Lagrange term for the sum of the probabilities q_j.

The EM algorithm is again an obvious method to estimate F, so we look at that option first. The self-consistency equation is given by:

$$\hat{q}_j = \frac{\hat{q}_j}{n(j+1)} \sum_{i=1}^{n} \frac{1}{\sum_{k \geq X_i} (1+k)^{-1} \hat{q}_k}\, 1_{\{j \geq X_i\}}\,, \tag{4.3}$$

where $\hat{q}_j = \hat{P}\{Y = j\}$ corresponds to the MLE of the mixing probabilities.

Note that (4.3 covers two situations

1. $\hat{q}_j > 0$, in which case the relation is equivalent to

$$\frac{1}{n(j+1)} \sum_{i=1}^{n} \frac{1}{\sum_{k \geq X_i} (1+k)^{-1} \hat{q}_k}\, 1_{\{j \geq X_i\}} = 1,$$

   and
2. $\hat{q}_j = 0$.

This is illustrated in Figure 3, where the mixing distribution if Poisson(10), and where the flat pieces of $\hat{F}_n$ correspond to sequences of probabilities $\hat{q}_i$ equal to zero. Note that the empirical distribution function is a much better estimate, but we cannot observe this function, due to the fact that we deal with an inverse

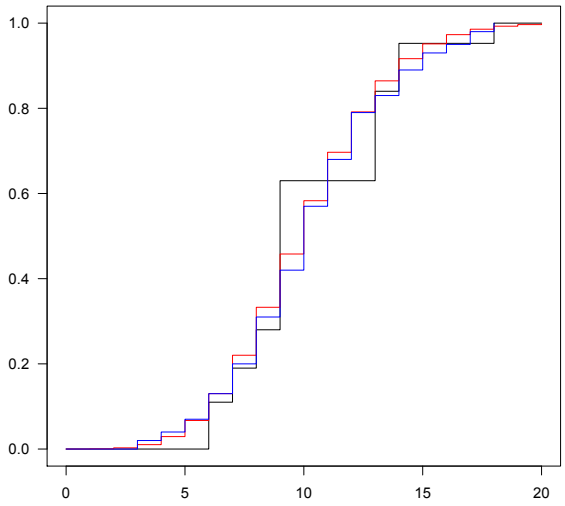

Fig 3: MLE $\hat{F}_n$ (black) of the mixing distribution function, empirical distribution function (blue) and real mixing distribution function (red), for a sample of size $n = 100$ from a Poisson distribution with parameter $\mu = 10$.

problem. We only can observe the observations of the *mixture* (instead of mixing) distribution in our sample! But if we generate the observations ourselves, we can of course plot this hidden empirical distribution function too.

Another method of computing the MLE of F is to use the fact that finding the MLE is equivalent to maximizing the log likelihood in the observation space over all decreasing discrete probability measures, under the restriction that the probabilities are decreasing in $i = 0, 1, \ldots$, which leads (analogously to the continuous case) to the discrete Grenander estimator. This is computed by constructing the least concave majorant of the cusum diagram, consisting of the point $(0, 0)$ and the points

$$(i, \mathbb{F}_n(i-1)), \qquad i = 1, 2, \ldots,$$

where $\mathbb{F}_n$ is the empirical distribution function of observed frequencies in the observation space.

Note the shift to the argument $i-1$ in the argument of $\mathbb{F}_n$, since counting starts at $i = 0$. For this reason [3] starts the cusum diagram with first coordinate -1 instead of 0. The cusum diagram consists of a finite number of points and ends at the index $m+1$, where m is the largest integer which has a positive frequency in the sample. The resulting decreasing density is shown in Figure 4.

The estimates of the mixture probabilities $p_i = P_F(X = i)$, $i = 0, 1, \ldots$ are now given by $\hat{p}_i = S_{i+1}$, where S_{i+1} is the left-continuous slope at $i+1$ of the cusum diagram, defined above, and where $\hat{p}_i = 0$, $i > m$. The MLE is then found from these estimates $\hat{p}_i$ by the relation

$$\hat{q}_i = \begin{cases} (i+1)\{\hat{p}_i - \hat{p}_{i+1}\}, & i = 0, \ldots, m, \\ 0, & i > m. \end{cases} \tag{4.4}$$

where $\hat{p}_i = 0$, $i > m$. Note that if $\hat{q}_i$ is defined in this way, we get:

$$\int_{y \geq j} \frac{1}{1+y}\, d\hat{F}(y) = \sum_{i \geq j} \frac{\hat{q}_i}{1+i} = \sum_{i=j}^{m} \{\hat{p}_i - \hat{p}_{i+1}\} = \hat{p}_j, \qquad 0 \leq j \leq m.$$

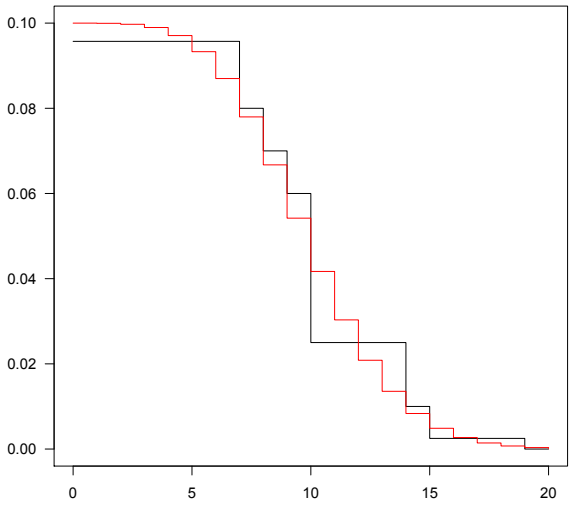

Fig 4: Grenander estimate (black) of the decreasing density P_F (red) of (4.2) for the same sample as used in Figure 3.

Since the $\hat{p}_j$'s are a sequence of nonincreasing numbers, the $\hat{q}_i$ are nonnegative, and

$$\sum_{i=0}^{m} \hat{q}_i = \sum_{i=0}^{m} (i+1)\,\{\hat{p}_i - \hat{p}_{i+1}\} = \sum_{i=0}^{m} \hat{p}_i = 1,$$

where the last equality holds since the $\hat{p}_i$ consist of partial means of the relative frequencies of the observations, which arise from the "pool adjacent violators" algorithm, and therefore sum to 1.

So the MLE of the mixing distribution can just be found by computing the Grenander estimator on the basis of the observations of the mixture distribution and using (4.4). Computing the Grenander estimator is a 1-step procedure and very much faster than the EM algorithm (which finds the same solution though, but usually needs more than 10^5 iterations to reach a solution with the same accuracy).

In contrast with the situation in the completely monotone case, the evaluation mapping $x \mapsto F(x)$ is differentiable. A boxplot of 1000 values of $\hat{F}_n(10)$ for the model (4.1), where $\mu = 10$, is shown in Figure 5 for $n = 100, 1000$ and $10,000$. The values of n times the variance increase somewhat: the values are 2.144767, 3.870744 and 4.06277, for $n = 100, 1000$ and $10,000$, respectively. But the values seem to stabilize around 4; for $n = 10^5$ we get 3.96917.

References

[1] Fadoua Balabdaoui and Gabriella de Fournas-Labrosse. Least squares estimation of a completely monotone pmf: From Analysis to Statistics. *J. Statist. Plann. Inference*, 204:55–71, 2020. ISSN 0378-3758. . URL https://doi.org/10.1016/j.jspi.2019.04.006.

[2] Piet Groeneboom and Geurt Jongbloed. *Nonparametric Estimation under Shape Constraints*. Cambridge Univ. Press, Cambridge, 2014.

[3] Hanna K. Jankowski and Jon A. Wellner. Estimation of a discrete monotone distribution. *Electron. J. Stat.*, 3:1567–1605, 2009. ISSN 1935-7524. . URL https://doi.org/10.1214/09-EJS526.

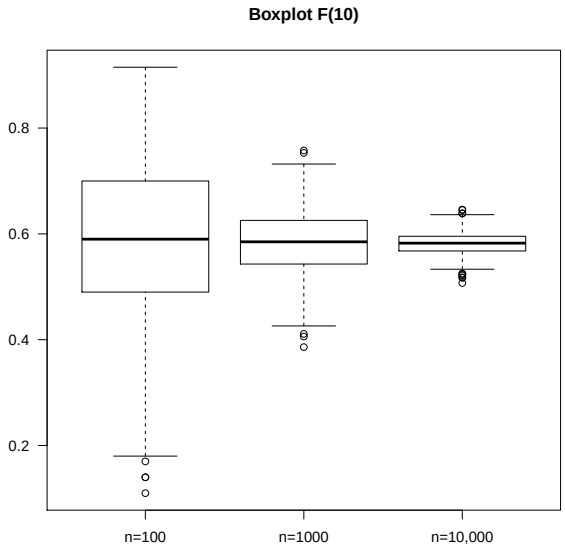

Fig 5: Boxplot of 1000 values of $\hat{F}_n(10)$ for the model (4.1), where $\hat{F}_n$ is the MLE of the mixing distribution, for sample sizes $n = 100, 1000$ and $10,000$.