CORRECTION ALGORITHM FOR WEIGHTED EMPIRICAL DISTRIBUTION FUNCTION

Kubaychuk O. O.
Candidate of Sciences in Physics and Mathematics, Associate Professor,
Professor at the Special Department № 1,
Institute of Specialized Communication and Information Security of the National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv, Ukraine

The main tool for studying mixtures with varying concentrations is the weighted empirical measures and weighted empirical distribution functions of the form

$$F_N^k(x) = \frac{1}{N} \sum_{j=1}^{N} a_j 1\{\xi_j < x\}$$

where $a_j$ are nonrandom fixed numbers.

The technique of a nonparametric analysis of mixtures where concentrations changes from observation to observation develops, actively. The problem of distributions estimating in case at known concentrations is considered in [1 p. 618], in particular, found conditions under which the weighted empirical distribution functions are unbiased and minimal estimators of unknown distribution functions of components of the mixture. The classification problem based on a sample of a mixture with varying concentrations is studied in [2, p. 342], [3, p. 47], [4, p. 78], [5, p. 6262].
Let $\Xi_N = (\xi_{1:N}, \ldots, \xi_{N:N})$ be a sample from a mixture with varying concentrations that is components are jointly independent random variables [6, p. 98], and

$$\Pr\{\xi_{j:N} < x\} = \sum_{m=1}^{M} w_{j:N}^m H_m(x),$$

where $M$ is the total number of components in the mixture, $H_m$ is the distribution function of the $m$-th component, and $w_{j:N}^m$ is the probability to observe an object from the $m$-th component in the $j$-th observation. Probability $w_{j:N}^m$ is called concentration or mixing probability. We assume that concentrations of components are known, and the distributions $H_m$ are unknown.

For example, the problem of estimation of functional moment for the $k$-th component of the mixture, namely

$$\bar{g}^k = \int g(x) H_k(dx),$$

where the real valued function $g$ is fixed, was studied in [7, p. 226] and [8, p. 7446]. One can consider the linear estimator for $\bar{g}^k$ becomes of the form

$$\hat{g}_N^k = \hat{g}_N^k(\bar{a}^k) = \int g(x) \hat{F}_N(dx, \bar{a}^k) = \frac{1}{N} \sum_{j=1}^{N} a_{j:N}^k g(\xi_{j:N}),$$

where

$$\hat{F}_N(x, \bar{a}) = \frac{1}{N} \sum_{j=1}^{N} a_{j:N} 1\{\xi_{j:N} < x\}, \quad (1)$$

are the weighted empirical distribution functions (wedf) with some nonrandom weight $\bar{a}$, are proposed in [1, p. 619] as estimators for $H_k$. It is shown in [1, p. 620] that the estimators,
defined (1), are unbiased, consistent, asymptotically normal and minimax for appropriate weight coefficients. However, it can be that the some coefficients \( a_{j:N} \) are negative, therefore the function \( \hat{F}_N(x, \bar{a}) \) can’t be a probability distribution function.

To improve the weighted empirical distribution functions \( \hat{F}_N(x, \bar{a}) \), put

\[
F_N^+(x, \bar{a}) = \sup_{y < x} \hat{F}_N(y, \bar{a}).
\]

(2)

The function \( F_N^+(x, \bar{a}) \) is nondecreasing and assumes only positive values, but it can assumes values greater then 1. Thus we consider the function

\[
\Phi_N^+(x, \bar{a}) = \min(1, F_N^+(x, \bar{a})).
\]

(3)

Accordingly, the estimators for functional moments becomes of the form

\[
\tilde{g}_N^{+k} = \int g(x)\Phi_N^+(dx, \bar{a}^k) = \frac{1}{N} \sum_{j=1}^{N} b_j^{+k} g(\xi_j),
\]

(4)

where \( b_j^{+k} \) are some coefficients that depend on the sample \( \Xi_N \). To obtain the improved distribution function \( \Phi_N^+(x, \bar{a}) \) (namely, coefficients \( b_j^{+k} \)) we can use the algorithm from [9, p. 95]. The complexity of the algorithm is \( O(N \ln N) \).

Similarly, we can construct an other estimators for functional moments [10, p. 48], [11, p. 83]. For example, put

\[
F_N^-(x, \bar{a}) = \inf_{y > x} \hat{F}_N(y, \bar{a}),
\]

(5)

\[
\Phi_N^-(x, \bar{a}) = \max(0, F_N^-(x, \bar{a})),
\]

(6)

and the combination of (3), (6), for example
\[ \Phi_N^\pm(x, \bar{a}) = \frac{1}{2}[\Phi_N^+(x, \bar{a}) + \Phi_N^-(x, \bar{a})]. \] \quad (7)

And, accordingly
\[ \tilde{g}_{N}^{-k} = \int g(x)\Phi_N^{-}(dx, \bar{a}_k) = \frac{1}{N} \sum_{j=1}^{N} b_{j}^{-k} g(\xi_j), \quad (8) \]

\[ \tilde{g}_{N}^{\pm k} = \int g(x)\Phi_N^{\pm}(dx, \bar{a}_k) = \frac{1}{N} \sum_{j=1}^{N} b_{j}^{\pm k} g(\xi_j), \quad (9) \]

where \( b_{j}^{-k} \), \( b_{j}^{\pm k} \) are some coefficients that depend on the sample \( \Xi_N \). To obtain the coefficients \( b_{j}^{-k} \), \( b_{j}^{\pm k} \) we can use the algorithms from [10, p. 50].

To evaluate the improved weighted empirical distribution functions (3), (6), (7), an effective algorithms were proposed in [9, p. 96], [10, p. 49], [13, p. 21] and [12, p. 315]. The asymptotic behavior was studied in [11, p. 84].

In what follows, we describe some algorithms of partially improving of weighted empirical distribution function. We assume, that members of the sample are arranged in ascending order and distinct. Note, that the number of operations required by fast sorting algorithm is \( O(N \ln N) \).

Let probabilities \( p_1 \) and \( p_2 \), \( 0 \leq p_1 < p_2 \leq 1 \) specify an interval of improving. The idea of the procedure is as follows. Moving from \( p_1 \) to \( p_2 \) (from left to right), we consecutively improve the coefficients \( a_j \), so that the sum \( \sum_{i: \xi_i \leq \xi_j} a_i \) become “upper” then all its predecessors. This algorithm is named as LR_UP(p1,p2).
The weighted empirical distribution functions are the powerful instrument of studying mixture with varying concentrations. To obtain the estimators that will be probability distribution functions the fast algorithm for the correction of weighted empirical distribution functions developed.

References:


8. Kubaychuk O. Functional moments estimators analysis by the Monte-Carlo method for model of mixture with varying concentrations.


