DOI https://doi.org/10.30525/978-9934-26-115-2-2

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

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

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

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

where

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

are the weighted empirical distribution functions (wedf) with some nonrandom weight $\vec{a}$, 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 $\widehat{F}_N(x,\vec{a})$ can't be a probability distribution function.

To improve the weighted empirical distribution functions $\widehat{F}_N(x,\vec{a})$, put

$$F_N^{+}(x,\vec{a}) \square \sup_{y<x} \widehat{F}_N(y,\vec{a}). \qquad (2)$$

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

$$\Phi_N^{+}(x,\vec{a}) = \min(1, F_N^{+}(x,\vec{a})) \ . \qquad (3)$$

Accordingly, the estimators for functional moments becomes of the form

$$\tilde{g}_N^{+k} = \int g(x)\Phi_N^{+}(dx,\vec{a}^{k}) = \frac{1}{N}\sum_{j=1}^{N} b_j^{+k} g(\xi_j), \quad (4)$$

where b_j^{+k} are some coefficients that depend on the sample Ξ_N. To obtain the improved distribution function $\Phi_N^{+}(x,\vec{a})$ (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

$$F_N^{-}(x,\vec{a}) \square \inf_{y>x} \widehat{F}_N(y,\vec{a}), \qquad (5)$$

$$\Phi_N^{-}(x,\vec{a}) = \max(0, F_N^{-}(x,\vec{a})) \ , \qquad (6)$$

and the combination of (3), (6), for example

$$\Phi_N^{\pm}(x,\vec{a}) = \frac{1}{2}[\Phi_N^{+}(x,\vec{a}) + \Phi_N^{-}(x,\vec{a})]. \qquad (7)$$

And, accordingly

$$\tilde{g}_N^{-k} = \int g(x)\Phi_N^{-}(dx,\vec{a}^k) = \frac{1}{N}\sum_{j=1}^{N} b_j^{-k} g(\xi_j), \qquad (8)$$

$$\tilde{g}_N^{\pm k} = \int g(x)\Phi_N^{\pm}(dx,\vec{a}^k) = \frac{1}{N}\sum_{j=1}^{N} b_j^{\pm k} g(\xi_j), \qquad (9)$$

where b_j^{-k}, $b_j^{\pm k}$ are some coefficients that depend on the sample Ξ_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, $0 \le p_1 < p_2 \le 1$ 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 $\sum_{i:\xi_i \le \xi_j} a_i$ 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:

1. Maiboroda R. E. Estimates for distributions of components of mixtures with varying concentrations. *Ukrainian Mathematical Journal*. 1996. *48*(4). P. 618-622.

2. Kubaychuk O. EBC-Estimator of Multidimensional Bayesian Threshold in Case of Two Classes. *Journal of Statistical Theory and Applications*. 2020. *19*(3). P. 342-351.

3. Кубайчук О. Асимптотика оцінки для баєсового порогу. *Вісник Київського національного університету імені Тараса Шевченка. Математика. Механіка*. 2008. 19-20. С. 47-50.

4. Kubaychuk O. O. The estimator asymptotic behavior of the empirical risk minimization method for bayesian border. *Research Bulletin of NTUU «Kyiv Polytechnic Institute»*. 2010. 4. P. 78-85.

5. Kubaychuk O. The asymptotic behaviour of threshold-based classification rules in case of three prescribed classes. *Journal of Advances in Mathematics*. 2016. *12*(5). P. 6262-6269.

6. Міхайленко В. М., Теренчук С. А., Кубайчук О. О. Теорія ймовірностей, ймовірнісні процеси та математична статистика. *К.: Вид-во Європ. ун-ту*. 2007. С. 121.

7. Kubaychuk O. Estimation of moments by observations from mixtures with varying concentrations. *Theory of Stochastic Processes*. 2002. Vol. 8, No. 3-4, P. 226-231.

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

concentrations. *Journal of Advances in Mathematics*. 2018. № 14(1). 7446–7451. https://doi.org/10.24297/jam.v14i1.6475

9. Maiboroda R., Kubaichuk O. Asymptotic normality of improved weighted empirical distribution functions. *Theory of Probability and Mathematical Statistics*. 2004. *69*. P. 95-102.

10. Kubaychuk O. O. Estimation of moments from mixtures with the use of improved weighted empirical distribution function. *Visnyk Kyiv Univ. Matematyka. Mekhanika*. 2003. *9*(10). P. 48-52.

11. Maiboroda R., Kubaichuk O. Improved estimators for moments constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics*. 2005. *70.* P. 83-92.

12. Kubaychuk O. Fast correction algorithms for weighted empirical distribution functions. *Advances and Applications in Statistics*. 2019. *54*(2). P. 315-326. http://dx.doi.org/10.17654/AS054020315

13. Kubaychuk O. O. Estimator of multidimensional Bayesian threshold in two-class classification // Topical issues and challenges of physical and mathematical sciences: conference proceedings. Wloclawek. Poland. 2021. P. 21–24. https://doi.org/10.30525/978-9934-26-043-8-5