A LOWER BOUND FOR THE MIXTURE PARAMETER AND ITS ESTIMATOR

Bobosharif K. Shokirov

bobosari@karlin.mff.cuni.cz, Bobosarif.Shokirov@chmi.cz

Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University in Prague

and

Department of Numerical Weather Prediction, Division of Meteorology and Climatology, Czech Hydrometeorological Institute

SUMMARY

This poster presents the problem of estimating the mixture parameter in two-component mixture model with one unknown component. Given a sample $X_1, \ldots, X_n$ of size n from a mixture $H(x;\theta)$ of two distribution functions $F(x)$ and $G(x)$, where $G(x)$ is unknown, an approach for estimating the mixture parameter θ is discussed. By utilizing the behavior of the family of random variables $\{\theta_n^*(x),\ x \in [0,\ 1],\ n = 1, 2, \ldots\}$, a number of properties of the estimator are derived. In particular, it is shown that this family of random variables contains an unbiased estimator of the mixture parameter. Based on the approach developed in [3], an inequality for the lower bound of the mixture parameter and its estimator is derived, which serves as the mixture parameter estimator.

THE SETTING OF THE PROBLEM

Let $X_1, \ldots, X_n$ be a sample of size n drawn from a distribution function (d.f.) $H(x;\theta)$ of the form

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad (\theta \in (0,1)). \tag{1}$$

In representation (1) $F(x)$ is a known d.f., while d.f. $G(x)$ and a parameter $\theta \in [0,1]$ are unknown. This is a binary mixture problem (or so called two-component mixture problem) and our aim is to estimate (non-parametrically) the parameter θ. Here notations $H(x)$, $H(x;\theta)$ and $H_\theta(x)$ are used interchangeably just to emphasize the that θ is a mixture parameter and d.f. $H(x)$ depends on θ through the mixture proportion.

Similar models can appear in many contexts. In multiple hypothesis testing problem when the number of hypotheses is large (for example in the analysis of differentially expressed genes, neuroimaging) the distribution of p-values under the null hypotheses as a result of independent statistical tests is uniform on the interval $[0,\ 1]$ (under continuity assumption), while under the alternative hypotheses is unknown (see, [1, 2]). In terms of the model (1), $F(x)$ is the uniform distribution (the distribution of p-values under the null hypotheses) and $G(x)$ is the distribution of the p-values under the alternative, which is unknown. In this context the aim is to estimate the proportion of false null hypotheses, that is, the mixture paramter θ.

In contamination problems under reasonable assumptions, distribution $F(x)$ can be contaminated by some arbitrary distribution $F_0(x)$, which yields a sample drawn from $H(x)$ as in (1) (see, for example, [5]). In astronomy, similar situations can arise quite often: once we observe a variable of interest (for example, metallicity, radial velocity) of stars in a distant galaxy, foreground stars from the Milky Way in the visible area, contaminate the sample. Stars in the galaxy can be difficult to distinguish from those of foreground stars since we are able only to observe the stereographic projections but not the 3D positions of the stars ([9]). Proceeding from the well-known physical models for the foreground stars, it is possible to constrain d.f. $F(x)$ and in addition to estimating the mixture parameter focus on estimating d.f. $F_0(x)$. High Energy Physics also can be a source of similar problems, where the evidence could be to have a significant peak at some position on top of some known distribution with nice properties (see, [4, 7]).

In contrast to the usual classical mixture problem, where the mixture consists of a combination of two or more, mainly specified or partially specified distributions, the mixture in the right-hand side of (1) contains an unknown component and hence suggested classical methods cannot be applied here. Instead, the approach proposed in [3] to a binary survival model seems to be more promising to drive an inequality for the mixture parameter and estimate its lower bound.

In the classical mixture problem (partially) specified components can already be considered as a certain type of restriction imposed on the family of distributions that together with other restrictions and assumptions makes the problem well-defined, in particular, identifiable. Basically, it is the identifiability of the model which makes it possible to estimate the mixture parameter. If we proceed from this principle, it seems impossible to estimate the mixture parameter without ensuring identifiability of the model. Although imposing certain restrictions makes model (1) well-defined enough, it still cannot ensure its identifiability.

However, it seems that even without ensuring identifiability, one can deal with the problem of estimating the mixture parameter θ in (1), in particular, one can derive certain bounds for it. To derive a lower bound for the mixture parameter θ in the model (1), without being specific, we impose certain restrictions on the components of the mixture model. First of all, we assume that

$$G(x) > F(x). \tag{2}$$

This assumption arises, for example, in multiple testing and certainly has practical application. Condition (2) still cannot guarantee the identifiability of the model, however it enables one to extract certain properties of the mixture parameter, in particular, it allows one to derive the lower bound inequality for the mixture parameter θ and obtain its estimator.

Without loss of generality we can assume that the support of d.f. $F(x)$ belongs to the interval $[0,\ 1]$ ($S_F = \mathrm{supp}F(x) \subset [0,\ 1]$), otherwise it could be transformed into the interval $[0,\ 1]$. Due to monotonicity, reducing the support S_F of d.f. $F(x)$ does not affect condition (2), that is, the inequality $G(x) > F(x)$ remains valid for $x \in [0,\ 1]$ and it guarantees that the support S_G of d.f. $G(x)$ be a proper subset of the support of d.f. $F(x)$: $S_G \subseteq S_F$. In general, the support S_G could be any proper subset of S_F of the forms $[0,\ 1-\delta]$, $[1-\delta,\ 1]$, $[\delta,\ 1-\delta]$ for some $0 < \delta < 1$. For now we assume that $S_G = [0,\ 1-\delta]$, $\delta > 0$. Other type of supports could be considered in a similar way. Therefore, it is enough to consider (1), defined on the interval $[0,\ 1]$.

Thus, after some assumptions and restrictions we arrive at the following problem

Estimate the mixture parameter θ in the model

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad x \in [0,\ 1], \quad (\theta \in (0,1)), \tag{3}$$

with the conditions

$$G(x) > F(x), \quad \forall x \in [0,\ 1] \tag{4}$$

and

$$S_G \subset [0,\ 1-\delta], \quad \text{for some} \quad \delta > 0. \tag{5}$$

In addition, we assume that d.f.'s $F(x)$ and $G(x)$ are continuously differentiable. It should be noted that among others, issues that bring to similar problem were considered in [6, 10].

PROPERTIES OF THE FAMILY OF R.V.'S $\{\theta_N^*(X),\ X \in [0,\ 1],\ N = 1, 2, \ldots\}$

Theorem 1 *Assume condition (4) holds. Let d.f.'s $F(x)$ and $G(x)$ are continuously differentiable and satisfy the relation*

$$\frac{F'(x)}{1-F(x)} \le \frac{G'(x)}{1-G(x)}. \tag{6}$$

Then the expected value of the family of random variables $\{\theta_n^*(x),\ x \in [0,1]\}$ is a monotonic non-increasing on the interval $[0,\ 1]$ function and $\theta \le \mathbb{E}[\theta_n^*(x)] \le 1,\ x \in [0,1]$.

For the proof of the theorem see [8].

Theorem 2 *Let conditions (4) and (5) be satisfied. Then if in addition to (6), the condition*

$$2\frac{F'(x)}{1-F(x)} \ge \frac{G'(x)}{1-G(x)} \tag{7}$$

also holds, then the variance $\sigma^2_{\theta_n^*(x)}$, defined as

$$\sigma^2_{\theta_n^*(x)} = \frac{H(x)(1-H(x))}{n(1-F(x))^2}. \tag{8}$$

is a monotonic nondecreasing function of x for all $x \in [0,1]$.

A LOWER BOUND ESTIMATOR

Lemma 1 .
Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be a sample of size n drawn from d.f. $H(x)$. Then sample $\mathbb{Y}_n = \{Y_1, \ldots, Y_n\}$ of size n drawn from the complementary cumulative distribution function (c.c.d.f.) $(1-H(x))/(1-F(x))$ could be obtained from $\mathbb{X}_n$ by

$$y = \overline{H}^{-1}\left(\frac{1-H(x)}{1-F(x)}\right), \quad \overline{H(x)} = 1 - H(x).$$

Proof. Since

$$\overline{H(x)} = \mathbb{P}\{X > x\},$$

hence for $x \in [0,\ 1]$

$$\frac{1-H(x)}{1-F(x)} = \frac{\mathbb{P}\{X > x\}}{1-F(x)} = \mathbb{P}\{Y > y\} = \overline{H(y)},$$

from which follows the statement of lemma.

Let us call $\mathbb{X}_n$ the original sample and $\mathbb{Y}_n$ its transformed sample.

Theorem 3 *Let $\mathbb{X}_n$ be the original sample and $\mathbb{Y}_n$ be its transformed sample and $1 \le k \le n$. Assume the following conditions hold:*

$$G(x) > F(x), \quad \forall x \in [0,\ 1], \tag{9}$$

$$S_G \subset [0,\ 1-\delta], \quad \text{for some} \quad \delta > 0, \tag{10}$$

and

$$\frac{F'(x)}{1-F(x)} \le \frac{G'(x)}{1-G(x)}. \tag{11}$$

Assume that $\varphi(x)$ is a strictly decreasing function on the interval $[0,\ 1]$ such that $\varphi(0) = -\varphi'(0) = 1$ and satisfies the relation

$$\frac{d^2}{dx^2}\left[\varphi^{-1}\left(\frac{1-H(x)}{1-F(x)}\right)\right] \ge 0. \tag{12}$$

Then for the mixture parameter in the model (3) the inequality

$$\theta \ge 1 - \frac{H(X) - F(X)}{\overline{F(X)}(1-\varphi(YR_H(y_0)))} \tag{13}$$

holds and the estimator of its lower bound, which serves as an estimator of the mixture parameter θ in the model (3), can be defined as

$$\theta_n^* = \max\left\{1 - \frac{k}{n[1-\varphi(YR_n(y_0))]}, 0\right\}, \tag{14}$$

where Y is defined as

$$\max\{Y_1, \ldots, Y_k\} \le Y \le \min\{Y_{k+1}, \ldots, Y_n\}, \quad k \le n, \tag{15}$$

$y_0 \in (0,\ Y)$, x_0 is such that $\overline{H(y_0)} \cdot \overline{F(x_0)} = \overline{H(x_0)}$ and

$$R_n(y_0) = \frac{1}{y_0}\varphi^{-1}\left(\frac{1-H_n(x_0)}{1-F(x_0)}\right),$$

$H_n(x)$ is the empirical d.f., constructed by the sample $\{X_1, \ldots, X_n\}$.

Proof of the theorem is based on the approach developed in [3], inequality (6), Lemma 1 and properties of the generalized hazard function $\varphi(x)$.

Acknowledgement. The author would like to express his thanks to Prof. Lev Klebanov for his generous support and valuable comments.

References

[1] Bar-Hen, A., Robin, S., Daudin, J.-J. and Pierre, L., A semi-parametric approach for mixture models: application to local false discovery rate estimation, *Comput. Statist. Data Anal*, —textbfVol. 51, 2007.

[2] Efron, B., Large-scale inference, *Institute of Mathematical Statistics Monographs*, **Vol. 1**, Cambridge University Press, Cambridge 2010.

[3] Klebanov, L. B. Yakovlev, A. A. Diverse correlation structures in gene expression data and their utility in improving statistical inference, *Statistics and Probability Letters*, **Vol. 31**, 2000.

[4] Lyons, L., Open statistical issues in particle physics, *Ann. Appl. Stat*, **Vol. 2**, 2008.

[5] McLachlan G., Peel, D,. Finite mixture models, Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York, 2000.

[6] Meinhausen, N., Rice, J. P. (2006) *Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses*, The Annals of Statistics, **Vol. 34**, 373–393.

[7] Patra, R., Sen, B.,Estimation of a two-component mixture model with applications to multiple testing, arXiv:1204.5488v1, 2012

[8] Shokirov, B. K. (2010) *On problem connected with the mixture parameter estimation*, Informační Bulleten Česke statistické společnosti, **Vol. 22**, 95–102.

[9] Walker, M.G., Mateo, M., Olszewski, E.W., Sen, B,. Woodroofe, M., Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present, *The Astronomical Journal*, **Vol. 137**, 2009.

[10] Wu, W. B. (2008) *On false discovery control under dependence*, The Annals of Statistics, **Vol. 36**, 364–380.