# A Lower Bound for the Mixture Parameter in the Binary Mixture Model and Its Estimator

Bobosharif K. Shokirov[1,2]

[1]Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague
and
[2]Department of Numerical Weather Prediction
Division of Meteorology and Climatology
Czech Hydrometeorological Institute

Robust 2012
10.09 – 14.09.2012, Němčičky

# Content

# The Setting of the Problem

Consider the following problem:

- Given a sample $X_1, \ldots, X_n$ of size $n$ from d.f. $H(x)$ of the form

$$H(x) = \theta F(x) + (1-\theta)G(x), \quad x \in \mathbb{R}, \quad (\theta \in (0,1)), \qquad (1)$$

# The Setting of the Problem

Consider the following problem:

- Given a sample $X_1, \ldots, X_n$ of size $n$ from d.f. $H(x)$ of the form

$$H(x) = \theta F(x) + (1-\theta)G(x), \quad x \in \mathbb{R}, \quad (\theta \in (0,1)), \qquad (1)$$

- where d.f. $F(x)$ is known but d.f. $G(x)$ and (mixture) parameter $\theta$ are unknown.

# The Setting of the Problem

Consider the following problem:

- Given a sample $X_1, \ldots, X_n$ of size $n$ from d.f. $H(x)$ of the form

$$H(x) = \theta F(x) + (1-\theta)G(x), \quad x \in \mathbb{R}, \quad (\theta \in (0,1)), \qquad (1)$$

- where d.f. $F(x)$ is known but d.f. $G(x)$ and (mixture) parameter $\theta$ are unknown.

- Aim: estimate $\theta$.

# Where this Model appears?

(i) Multiple hypotheses testing procedure.

# Where this Model appears?

(i) Multiple hypotheses testing procedure.

(ii) In astronomy (contamination of d.f. $F(x)$ )

# Where this Model appears?

(i) Multiple hypotheses testing procedure.

(ii) In astronomy (contamination of d.f. $F(x)$ )

# Where this Model appears?

(i) Multiple hypotheses testing procedure.

(ii) In astronomy (contamination of d.f. $F(x)$ )

*Under reasonable assumptions, d.f. $F(x)$ can be contaminated by some d.f. $F_0(x)$, which yields a sample from $H(x)$ as in (1) (see, for example, [3]). 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 ([5]). Due to physical models for the foreground stars, one can constrain d.f. $F(x)$ and focus on estimating the mixture parameter (and d.f. $F_0(x)$). High Energy Physics also can be a source of similar problems, where the evidence could a significant peak at some position on top of some known distribution with nice properties (see, [2]).*

(iii) etc.

# Identifiability

- Model (1) is not identifiable!

Why?

- Because with the same d.f. $F(x)$ one can construct another representation, different from (1).

Then why this Model?

# Identifiability

- Model (1) is not identifiable!

Why?

- Because with the same d.f. $F(x)$ one can construct another representation, different from (1).

Then why this Model?

- Because in this model one can estimate $\theta$!

# Assumptions

Assume

(i)

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

# Assumptions

Assume

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

(ii) The support $S_F$ of d.f. $F(x)$ belongs to the interval $[0,\ 1]$,

# Assumptions

Assume

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

(ii) The support $S_F$ of d.f. $F(x)$ belongs to the interval $[0,\ 1]$,

(iii)
$$\frac{F'(x)}{1 - F(x)} \leq \frac{G'(x)}{1 - G(x)}. \tag{3}$$

# Assumptions

Assume

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

(ii) The support $S_F$ of d.f. $F(x)$ belongs to the interval $[0,\ 1]$,

(iii)
$$\frac{F'(x)}{1 - F(x)} \leq \frac{G'(x)}{1 - G(x)}. \tag{3}$$

# Assumptions

Assume

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

(ii) The support $S_F$ of d.f. $F(x)$ belongs to the interval $[0,\ 1]$,

(iii)
$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}. \tag{3}$$

Then due to monotonicity of d.f.'s (2) remains valid and $S_G \subseteq S_F$.

# The Setting of the Problem after transformation

Estimate parameter $\theta$ in the model

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad x \in [0,\ 1], \quad (\theta \in (0,1)), \qquad (4)$$

# The Setting of the Problem after transformation

Estimate parameter $\theta$ in the model

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad x \in [0,\ 1], \quad (\theta \in (0,1)), \tag{4}$$

with conditions

(A1)

$$G(x) > F(x), \quad \forall x \in [0,1], \tag{5}$$

where d.f. $F(x)$ is known, while d.f. $G(x)$ is unknown, $x \in [0,\ 1]$.

# The Setting of the Problem after transformation

Estimate parameter $\theta$ in the model

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad x \in [0,\ 1], \quad (\theta \in (0,1)), \tag{4}$$

with conditions

(A1)

$$G(x) > F(x), \quad \forall x \in [0,1], \tag{5}$$

(A2)

$$\text{supp} G(x) \subset [0, 1-\delta], \quad \text{for some} \quad \delta > 0, \tag{6}$$

where d.f. $F(x)$ is known, while d.f. $G(x)$ is unknown, $x \in [0,\ 1]$.

# The Setting of the Problem after transformation

Estimate parameter $\theta$ in the model

$$H(x;\theta) = \theta F(x) + (1-\theta)G(x), \quad x \in [0,\ 1], \quad (\theta \in (0,1)), \tag{4}$$

with conditions

(A1)

$$G(x) > F(x), \quad \forall x \in [0,1], \tag{5}$$

(A2)

$$\text{supp}G(x) \subset [0, 1-\delta], \quad \text{for some} \quad \delta > 0, \tag{6}$$

(A3)

$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}. \tag{7}$$

where d.f. $F(x)$ is known, while d.f. $G(x)$ is unknown, $x \in [0,\ 1]$.

# Transformation of the Sample

**Lemma**

*Let $\mathbb{X}_n = \{X_1, \dots, X_n\}$ be a sample of size n drawn from d.f. $H(x)$. Then sample $\mathbb{Y}_n = \{Y_1, \dots, Y_n\}$ 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*

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

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

# A Lower Bound and its Estimator

## Theorem

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

$$G(x) > F(x), \quad \forall x \in [0,\ 1], \tag{8}$$

$$S_G \subset [0,\ 1-\delta], \quad \textit{for some} \quad \delta > 0, \tag{9}$$

*and*

$$\frac{F'(x)}{1-F(x)} \leq \frac{G'(x)}{1-G(x)}. \tag{10}$$

# A Lower Bound and its Estimator (cont.)

**Theorem (cont.)**

*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*

$$\frac{d^2}{dx^2}\left[\varphi^{-1}\left(\frac{1-H(x)}{1-F(x)}\right)\right] \geq 0. \tag{11}$$

*Then for the mixture parameter in the model (4) the inequality*

$$\theta \geq 1 - \frac{H(X) - F(X)}{\overline{F(X)}(1-\varphi(YR_H(y_0)))} \tag{12}$$

*holds and the estimator of its lower bound, which serves as an estimator of $\theta$ in the model (4), can be expressed as*

# A Lower Bound and its Estimator (cont.)

## Theorem (cont.)

$$\theta_n^* = \max\left\{1 - \frac{k}{n[1-\varphi(YR_n(y_0))]}, 0\right\}, \tag{13}$$

*where $Y$ is defined as*

$$\max\{Y_1, \ldots, Y_k\} \le Y \le \min\{Y_{k+1}, \ldots, Y_n\}, \quad k \le n, \tag{14}$$

*$y_0 \in (0,\ Y)$, $x_0$ 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 $\{X_1, \ldots, X_n\}$.*

# Thank You!

# References

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

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

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

Robin, S., Bar-Hen, A,. 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.

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.