Finite mixtures analysis by biased samples

Olena Sugakova[1] and Rostyslav Maiboroda[2]

[1]Kyiv National University, Ukraine, e-mail: sugak@univ.kiev.ua
[2]Kyiv National University, Ukraine, e-mail: mre@univ.kiev.ua

Abstract

We propose new estimates for means and CDFs of finite mixture components when the mixing proportions are not constants and some samping bias is present.

Keywords: Horwitz-Thompson sampling bias correction, weighted empirical cumulative distribution function, mixture with varying concentrations, finite mixture model

1 Introduction

In this presentation we discuss means and CDFs estimation of finite mixture components in the case when the observed sample is subjected to some sampling bias. The proposed estimation technique borrows from the Horwitz-Thompson (HT, see Lohr, 2010, p. 241) bias correction and the methodology of mixtures with varying concentrations (MVC) analysis, see Maiboroda & Sugakova (2012). In what follows we present a motivating example (Section 2), describe the HT estimates by a homogeneous sample (Section 3) and the MVC estimates by unbiased samples (Section 4). Then the estimates by both mixed and biased data are introduced in Section 5. Results of simulations are presented in Section 6.

2 Motivating example

Imagine that the distribution of some characteristic ξ (e.g. body length) of crabs living at a sea is studied. The investigated population of crabs is divided into two sub-populations (components). The crabs belonging to the first component are more salt-loving then the ones belonging to the second component. Both components are present in all sites in the sea where the crabs are caught, but their proportion in the local population depends on the mean salinity of the water at this site. Assume that the function describing the dependence of this proportion from the salinity is known.

We are interested in the differences in distribution of ξ for crabs belonging to different components. But the true component to which the crab belongs is not observed in the study, since it needs some expensive and time consuming tests. Therefore our inference should be based on the proportions of components at the sites where the crabs were caught. These proportions can be considered as probabilities that a crab chosen at random from the local population belongs to a given component (mixing probabilities). They can be estimated by the mean salinity data.

To catch the crabs some traps are used and it is known that the probabilities to be caught are different for crabs with different body length ξ. This causes a sampling bias in the observed distribution of ξ. Our aim is to correct this bias and to extract the CDF of the component of interest from the mixture.

3 Horwitz-Thompson approach to bias correction

Assume that there is a homogeneous infinite (very large) population of subjects O with the observed feature $\xi(O) \in \mathbb{R}$. The CDF of $\xi(O)$ in the entire population is $F(x)$. A subject O can be sampled from the population with the probability depending on $\xi(O)$ but independently of all other subjects. (See Shao, 2003, p. 328). Let us denote this (inclusion) probability by

$$cq(t) = \mathsf{P}\{O \text{ was included to the sample } |\xi(O) = t\}, \tag{1}$$

where $q(t)$ is a known function, c is an unknown constant. The values of $\xi(O)$ for the sampled subjects constitute the sample $Y = (\eta_1, \eta_2, \ldots, \eta_n)$. It is obvious that η_i are i.i.d. and the their CDF $\tilde{F}(x)$ is the conditional probability of the event $\{\xi(O) < x\}$ given that O was sampled. Then

$$\tilde{F}(x) = \mathsf{P}\{\xi(O) < x \mid O \text{ was sampled}\} = \frac{\int_{-\infty}^{x} q(t)F(dt)}{\int_{-\infty}^{+\infty} q(t)F(dt)}. \tag{2}$$

In this case the population mean $\bar{\xi} = \mathsf{E}\,\xi = \int xF(dx)$ does not equal to the expectation of the observed values $\mathsf{E}\,\eta_j$ due to the sampling bias. But $\bar{\xi}$ can be estimated by the weighted sample mean with weights reciprocal to the inclusion probabilities:

$$\hat{\xi} = \frac{1}{\sum_{j=1}^{n} \frac{1}{q(\eta_j)}} \sum_{j=1}^{n} \frac{1}{q(\eta_j)} \eta_j.$$

It is the usual HT-estimate which is known to be consistent if $\bar{\xi}$ exists and $q(t) > \mathit{const} > 0$ for all t. The corresponding estimate for CDF F is

$$\hat{F}^{HT}(x) = \frac{1}{\sum_{j=1}^{n} \frac{1}{q(\eta_j)}} \sum_{j=1}^{n} \frac{1}{q(\eta_j)} \mathbb{1}\{\eta_j < x\}.$$

4 Mixtures with varying concentrations

In the MVC model we assume that the subjects can belong to one of M sub-populations (components) $\mathcal{P}_1$, $\mathcal{P}_2$,…,$\mathcal{P}_M$. The probability to observe a subject from a given component depends on the conditions of the observations and is, generally speaking, different for different observations. Let us denote by p_j^i the probability to observe a subject from $\mathcal{P}_i$ in the j-th observation. The CDF of the observed feature $\xi(O)$ of a subject O is different for different components: $H_m(x) = \mathsf{P}\{\xi(O) < x \mid O \in \mathcal{P}_m\}$. So, the observed sample X consists of independent but not identically distributed observations $X = (\xi_1, \ldots, \xi_n)$ with CDFs

$$F_j(x) = \mathsf{P}\{\xi_j < x\} = \sum_{m=1}^{M} p_j^m H_m(x). \tag{3}$$

Note that the component to which an observed subject belongs is unknown. One needs to infer on H_m only by the sample X and the set of mixing probabilities (concentrations) p_j^m, $j = 1, \ldots, n$, $m = 1, \ldots, M$ which are known (estimated). We do not assume any sampling bias now.

The CDF H_m may be estimated by a weighted empirical CDF

$$\hat{H}_m(x) = \frac{1}{n} \sum_{j=1}^{n} a_j^m \mathbb{1}\{\xi_j < x\}.$$

Here a_j^m are some weights which may depend on mixing probabilities p_j^i, but not on the observations ξ_j.

To obtain unbiased estimates one needs the following conditions to be satisfied:

$$\frac{1}{n} \sum_{j=1}^{n} a_j^m p_j^i = \mathbb{1}\{i = m\} \text{ for all } i = 1, \ldots, M.$$

One of appropriate choices is the minimax weighting with

$$a_j^m = \sum_{i=1}^M \bar{\gamma}_{im} p_j^i,$$

where $\bar{\Gamma} = (\bar{\gamma}_{im})_{i,m=1}^M$ is the matrix inverse to $\Gamma = (\frac{1}{n}\sum_{j=1}^n p_j^i p_j^k)_{i,k=1}^M$.

To estimate the m-th component mean $\bar{\xi}_m = \int x H_m(dx)$ one may use

$$\hat{\xi}_m = \frac{1}{n}\sum_{j=1}^n a_j^m \xi_j.$$

5 Mixture model with sampling bias

Let us assume now that the MVC model is in force together with the sampling bias. It means that the considered subjects O belong to M different components and the CDF of their feature of interest $\xi(O)$ is H_m for the subjects belonging to the m-th component. The proportion of the m-th component subjects in the local population from which the j-th subject was obtained is p_j^m. These probabilities are known. The CDFs H_m are unknown. The sampling is biased in the sense that the probability to sample the subject O from a local population depends on $\xi(O)$. This probability is defined by (1) with known q and unknown c.

The problem is to estimate the components' CDFs H_m and means $\bar{\xi}_m = \int x H_m(dx)$.

Analogously to (2) we obtain that the observed sample $Y = (\eta_1, \dots, \eta_n)$ consists of independent observations with CDFs

$$\tilde{F}_j(x) = \mathsf{P}\{\eta_j < x\} = \frac{\int_{-\infty}^x q(t) F_j(dx)}{\int_{-\infty}^{+\infty} q(t) F_j(dx)}, \tag{4}$$

where F_j is defined by (3).

From (4) we obtain

$$\tilde{F}_j(x) = \sum_{m=1}^M \frac{p_j^m \tilde{Q}_m}{\sum_{i=1}^M p_j^i \tilde{Q}_i} \tilde{H}_m(x),$$

where $\tilde{Q}_m = \int_{-\infty}^{\infty} q(t) H_m(dx)$, $\tilde{H}_m(x) = \int_{-\infty}^{x} q(t) H_m(dx)/\tilde{Q}_m$. So, the sampling bias causes changes not only in the distributions of components, but also in the mixing probabilities. To take in account the bias in the mixing probabilities, we need to estimate $\tilde{Q}_m$. This can be done by observing that

$$\mathsf{E}\frac{1}{q(\eta_j)} = \frac{1}{\sum_{i=1}^M p_j^i \tilde{Q}_m}.$$

Then the least squares technique suggests the estimate $\hat{\mathbf{Q}} = (\hat{Q}_1, \dots, \hat{Q}_M)$ for $\tilde{\mathbf{Q}} = (\tilde{Q}_1, \dots, \tilde{Q}_M)$ which is the minimizer of the LS functional

$$J(\mathbf{Q}) = \sum_{j=1}^n \left(\frac{1}{\sum_{i=1}^M p_j^i Q_i} - \frac{1}{q(\eta_j)} \right)^2$$

over all $\mathbf{Q} = (Q_1, \dots, Q_M)$ with $Q_i > 0$.

With these estimates at hands we define the weights $\tilde{a}_j^m$ for the m-th component estimation as

$$\tilde{a}_j^m = \frac{1}{q(\eta_j)} \sum_{k=1}^M \tilde{\gamma}_{km} \frac{p_j^k}{\sum_{i=1}^M p_j^i \hat{Q}_i},$$

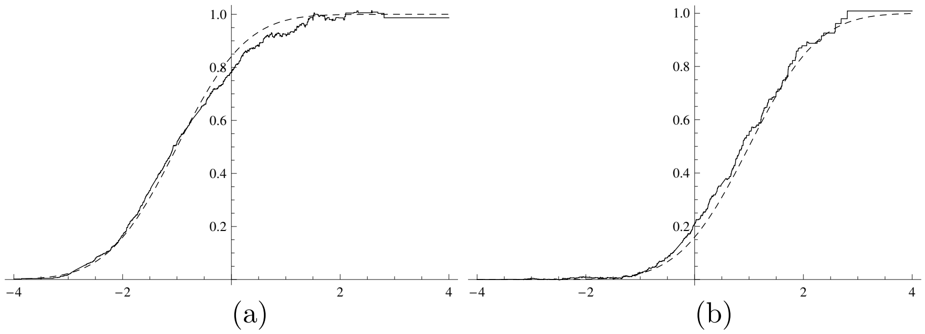

Figure 1: Estimates of CDF for the first (a) and second (b) component. The true CDFs are depicted by dashed lines.

where $\hat{\Gamma}_Q = (\hat{\gamma}_{km})_{k,m=1}^{M}$ is the matrix inverse to

$$\tilde{\Gamma}_Q = \left(\frac{1}{n} \sum_{j=1}^{n} \frac{p_j^k p_j^m}{\left(\sum_{i=1}^{M} p_j^i \hat{Q}_i \right)^2} \right)_{k,m=1}^{M}.$$

The resulting bias-correcting estimate for the H_m is

$$\hat{H}_m^{BC}(x) = \frac{1}{n} \sum_{j=1}^{n} \tilde{a}_j^m \mathbb{1}\{\eta_j < x\}.$$

The estimate for $\bar{\xi}_m$ is

$$\hat{\xi}_m^{BC} = \frac{1}{n} \sum_{j=1}^{n} \tilde{a}_j^m \eta_j.$$

6 Results of simulation

We performed a small simulation study to assess performance of the proposed estimates. In our experiment we considered a two component mixture $M = 2$. The distribution of the first component was $N(-1, 1)$, the distribution of the second one was $N(1, 1)$. The concentrations of the first component p_j^1 were simulated as random variables, uniformly distributed on $[0, 1]$, $p_j^2 = 1 - p_j^1$. Figure 1 presents the graphs of the estimates $\hat{H}_m^{BC}(x)$ for the components CDFs by a sample with $n = 1000$ observations.

The biases and variances of the estimates $\hat{\xi}_m^{BC}$ for different sample sizes n are presented in Table 1.

Table 1: Performance of the estimates for means

n	ξ_1^{BC}		ξ_2^{BC}	
	bias	Var	bias	Var
50	-0.0807	0.8011	-0.0851	0.3360
100	-0.0581	0.2257	-0.0589	0.1575
250	-0.0176	0.0849	-0.0271	0.0815
500	-0.045	0.0464	0.00210	0.0330
750	-0.0285	0.0421	-0.0162	0.0187
1000	-0.0052	0.0211	-0.0034	0.0118

7 Concluding remarks

We have constructed the estimates for means and CDFs of mixture components in the case when the sampling procedure is biased. The simulations indicate satisfactory behavior of the estimates in the case of Gaussian mixture. More efforts are needed to analyze the asymptotic behavior of these estimates and their performance on non-Gaussian data.

References

Lohr, S. (2010) *Sampling: Design and Analysis*. Brooks/Cole.

Maiboroda, R. & Sugakova, O.(2012) *Statistics of mixtures with varying concentrations with application to DNA microarray data analysis*. Nonparametric statistics, **24**, iss.1, 201-215.

Shao, J. (2003) *Mathematical statistics*. Shpringer.