TECHNICAL REPORT 0106U005864-MS11

Statistics of mixtures with varying concentrations with application to DNA microarray data analysis

Rostyslav Maiboroda (Kyiv National University, Ukraine)
Olena Sugakova (Kyiv National University, Ukraine)
e-mail: mre@univ.kiev.ua

Abstract

A finite mixture model is considered in which the mixing probabilities vary from observation to observation. Estimation of mixture components distributions, functional moments and densities is discussed. Tests are proposed for testing hypotheses on the moments. An application to the analysis of DNA microarrays data is considered.

Key words: finite mixture model; homogeneity test; density estimation; generalized moments estimation; weighted empirical distribution function; minimax weights

Kyiv - 15.02.2011

1. Introduction

In this paper nonparametric estimation and hypotheses testing problems are considered for observations described by the model of finite mixture with varying concentrations (MVC). In classical finite mixture models (FMM) one deals with i.i.d. observations of variables describing some subjects, belonging to M components (subpopulations) of the mixture. The true component to which a subject belongs is unknown, so the distribution of an observed variable ξ is

$$\mathsf{P}\{\xi \in A\} = p^1 F_1(A) + p^2 F_2(A) \cdots + p^M F_M(A),$$

where F_m is the distribution of ξ for subjects from the m-th component of the mixture, p^m is the probability to observe a subject from the m-th component (mixing probability, concentration of the component in the mixture). See McLachlan and Peel (2000), Titterington et al. (1985).

In the MVC model the observed variables $\xi_1,\ldots,\xi_n$ are also independent but not identically distributed since the mixing probabilities p_j^m vary from observation to observation:

$$\mathsf{P}\{\xi_j \in A\} = p_j^1 F_1(A) + p_j^2 F_2(A) \cdots + p_j^M F_M(A). \tag{1}$$

This model is a natural generalization of the multisample finite mixture model with different mixing probabilities in different samples (see Titterington et al., 1985). Its applications to the analysis of medical and biological data are discussed in Lodatko and Maiboroda (2006). Another application to genetical studies data is presented below in Sections 2 and 5.2. Applications in social science, finance and image analysis are considered in Autin and Pouet (2010).

The first problem which arises in the analysis of mixture models is their identifiability. There are many results on consistent estimation in parametric FMMs (e.g. Holzmann et al., 2006 and references cited therein). Identifiability issues for some nonparametric FMMs were discussed in Hall and Zhou (2003), Hunter et al. (2007), Bordes et al. (2006), Maiboroda (2008). In the MVC case there is a simple condition (13) on mixing probabilities p_j^m which assures identifiability of all components distributions F_m in nonparametric setting. In this paper we discuss a more complicated case when some of F_m are identifiable but some others are not. We assume here that the mixing probabilities are known, but no assumptions on the components distributions are made. The aim is to estimate F_m or its characteristics such as moments or densities and to test hypotheses on F_m.

The rest of the paper is organized as follows. In Section 2 we discuss possible applications of MVC to the DNA microarray data analysis. Some general definitions and notations are given in Section 3. Main statistical algorithms and their asymptotic properties are discussed in Section 4. Results of simulations and an application to a real DNA microarray data are presented in Section 5. Details of proofs are placed in Appendix.

2. Mixture analysis for DNA microarray data

Let us consider possible application of mixture models to the analysis of DNA microarrays data. DNA microarrays are a modern technology of genetical studies in which the RNA obtained from an investigated tissue (target RNA) is compared to a set of known DNA sequences called probes. The probes usually represent known or unknown genes.

In this technique cDNA obtained by reverse transcription of the target RNA is hybridized with the probe oligonucleotides, fixed in different spots on a silicon chip. The probes and targets are in different colors (say, green probes and red targets) so the hybridization changes the color of the spot. The change of color is automatically detected by a laser scanner. As a result the proportion of target complementary to each probe is derived. It represents the level of probe genes expression in the investigated tissue.

One of the main goals of such data analysis is to determine differences in genes expression in different tissues or under different conditions. Let there be n genes O_1, $\ldots$, O_n for which the expression levels were measured in two types of tissues (say, a normal one and a malignant one). For each gene O_j some statistics $\tau_j = \tau(O_j)$ is calculated over the data on its expression (E.g. it may be the Student-T statistics for the hypotheses of means homogeneity for expression levels of O_j in both tissues). This statistics has high values if O_j is differently expressed in tissues of different types and low values if O_j expression does not differ in these tissues.

Biologists compare τ_j to some fixed threshold T and conclude that O_j is a differently expressed, "interesting" gene if $\tau_j > T$. If $\tau_j < T$ then O_j is considered as an "uninteresting" gene (not involved in the difference between the investigated tissue types). Then the data on interesting and uninteresting genes can be analyzed separately as two different samples. But, in fact, for many practical test procedures no level of T can assure absence of errors in this classification procedure. Moreover, the probability of the error can be so large that the scientist can not recognize any gene as interesting with a satisfactory level of certainty.

In this case it is natural to consider the set of observations τ_j as taken from the mixture of two components: the first component ($m = 1$) corresponds to the interesting genes, the second one ($m = 2$) corresponds to the uninteresting ones. The distribution of τ_j is then described by the model

$$\mathsf{P}\{\tau_j < t\} = qF_1^\tau(t) + (1-q)F_2^\tau(t),$$

where F_m^τ, $m = 1, 2$ are the CDFs of τ for corresponding components and q is, roughly speaking, the proportion of interesting genes in the data.

When no additional assumptions are made on F_m^τ this model is unidentifiable. But if one assumes a parametric model for F_m^τ (say, Gaussian) it is possible to estimate the components distributions and the mixing proportion q, see Titterington et al. (1985), Tanaka (2009). Moreover, a consistent estimation is possible in a nonparametric setting when one component is assumed to be symmetrically distributed and the distribution of the other one is known, see Bordes at al (2006a), Maiboroda and Sugakova (2010).

With q and F_m^τ at hands the posterior probability

$$\tilde{p}(t) = \mathsf{P}\{O \text{ belongs to the first component} \mid \tau(O) = t\}$$

can be evaluated. Say, if $F_m^\tau \sim N(a_m, \sigma^2)$ then $\tilde{p}(t) = 1/(1 + \exp(\gamma_0 + \gamma_1 t))$ where $\gamma_0 = \log((1-q)/q) + (a_1^2 - a_2^2)/2$, $\gamma_1 = a_2 - a_1$.

So, for any gene O_j we can't determine with confidence to what component it belongs, but can estimate the probability $p_j^1 = p(\tau_j)$ that it belongs to the first component.

Now let there be data on the genes O_j expression levels $\xi_j = \xi(O_j)$ in a tissue of some third type (say an embryonal one). We would like to know how $\xi(O)$ is distributed for interesting and uninteresting genes O in this tissue. Denoting these

distributions as F_m^ξ we obtain the following model for the distribution of ξ_j:

$$\mathsf{P}\{\xi_j \in A\} = p_j^1 F_1^\xi(A) + (1 - p_j^1) F_2^\xi(A)$$

(conditionally on τ_j).

This is just the MVC model of the form (1) with two components. In what follows we discuss how to estimate the distributions F_m^ξ and to test hypotheses on their differences. Say, is there a difference in mean expression of interesting and uninteresting genes in the considered tissue?

3. Notations and definitions

Let the observed variable ξ be a random element of some measurable space $\mathcal{X}$ with the σ-algebra of measurable sets $\mathfrak{A}$. For asymptotic analysis we consider the observed sample $\Xi_n = (\xi_{1:n}, \dots, \xi_{n:n})$ as an element of an (imaginary) series $\Xi_1, \dots, \Xi_n, \dots$ assuming that $\xi_{j:n}$ are independent for fixed n and, for all $A \in \mathfrak{A}$,

$$\mathsf{P}\{\xi_{j:n} \in A\} = \Psi_{j:n}(A) := \sum_{m=1}^{M} p_{j:n}^m F_m(A), \tag{2}$$

where F_m are some probabilistic measures on $(\mathcal{X}, \mathfrak{A})$, $p_{j:n}^m$ are some real numbers satisfying the assumptions $p_{j:n}^m \geq 0$, $\sum_{j=1}^n p_{j:n}^m = 1$ for all $m = 1, \dots, M$, $j = 1, \dots, n$, $n = 1, 2, \dots$

To simplify notations when n is fixed we drop the subscript $_{:n}$, so $\xi_j = \xi_{j:n}$, $p_j^m = p_{j:n}^m$ and so on.

For the array $\mathbf{p} = (p_{j:n}^m,\ j = 1, \dots, n,\ m = 1, \dots, M,\ n = 1, 2, \dots)$ the symbol $\mathbf{p}_{:n}^m$ means the vector $(p_{1:n}^m, \dots, p_{n:n}^m)^T$, $\mathbf{p}_{j:n}$ means $(p_{j:n}^1, \dots, p_{j:n}^M)^T$ and $\mathbf{p}_{:n}$ means the $n \times m$ matrix $(p_{j:n}^m,\ j = 1, \dots, n,\ m = 1, \dots, M)$. The same notation is used for any other array of analogous structure, e.g. for $\mathbf{a} = (a_{j:n}^m,\ j = 1, \dots, n,\ m = 1, \dots, M,\ n = 1, 2, \dots)$.

The angle brackets denote the operator of averaging over j:

$$\langle \mathbf{a}^m \rangle_n := \frac{1}{n} \sum_{j=1}^{n} a_{j:n}^m.$$

Multiplication, summation and other arithmetic operations are applied to the arrays elementwise, so

$$\langle \mathbf{a}^m, \mathbf{p}^i \rangle_n := \langle \mathbf{a}^m \mathbf{p}^i \rangle_n = \frac{1}{n} \sum_{j=1}^{n} a_{j:n}^m p_{j:n}^m,\ \langle (\mathbf{a}^m)^2 \rangle_n = \frac{1}{n} \sum_{j=1}^{n} \left(a_{j:n}^m\right)^2.$$

Define $\langle \mathbf{a}^m \rangle := \lim_{n\to\infty} \langle \mathbf{a}^m \rangle_n$ if this limit exists. Note that $\langle \mathbf{a}^m, \mathbf{p}^i \rangle_n$ is an inner product on $\mathbb{R}^n$ and $\langle \mathbf{a}^m \mathbf{p}^i \rangle = \langle \mathbf{a}^m, \mathbf{p}^i \rangle$ may be considered as an inner product on the space of "triangular" arrays such as $\mathbf{a}^m$ or $\mathbf{p}^m$.

To clarify notation we introduce formally random variables η_m, $m = 1, \dots, M$ with distributions F_m. So for any function $\mathbf{g} : \mathcal{X} \to \mathbb{R}^d$,

$$\mathsf{E}\, \mathbf{g}(\xi_{j:n}) = \sum_{m=1}^{M} p_{j:n}^m\, \mathsf{E}\, \mathbf{g}(\eta_m).$$

The symbol $\Rightarrow$ means weak convergence of distributions and $\mathbb{1}(A)$ means the indicator function of a set A.

4. Estimation and testing

4.1. **Weighted empirical measures.** Further we consider the mixing probabilities $p^m_{j:n}$ as known and F_m as unknown. On the estimation of $p^m_{j:n}$ see Maiboroda (1993,2002).

Let us start with the estimation of the distributions $F_i(\cdot)$ as probability measures on $\mathcal{X}$. We will use the weighted empirical measures

$$\hat{F}_{\mathbf{a}:n}(A) = \hat{F}_{\mathbf{a}}(A) := \frac{1}{n}\sum_{j=1}^{n} a_j \mathbb{1}\{\xi_j \in A\} \tag{3}$$

as estimates for $F_m(A)$. Here $\mathbf{a} = (a_j, j = 1, \dots, n)$ are some nonrandom weights (dependent on $\mathbf{p}_{\cdot n}$) aimed to set off the m-th component and to suppress the influence of all other components in the sample.

To be an unbiased estimate for $F_i(A)$, $\hat{F}_{\mathbf{a}}(A)$ must satisfy

$$\mathsf{E}_{\mathbf{F}}\, \hat{F}_{\mathbf{a}}(A) = \frac{1}{n}\sum_{j=1}^{n} a_j\, \mathsf{E}_{\mathbf{F}}\, \mathbb{1}\{\xi_j \in A\} = \sum_{m=1}^{m} \langle \mathbf{a}\mathbf{p}^m\rangle_n F_m(A) = F_i(A)$$

for all possible sets of distributions $\mathbf{F} = (F_1(\cdot), \dots, F_M(\cdot))$. So the unbiasedness of $\hat{F}_{\mathbf{a}}$ is equivalent to

$$\langle \mathbf{a}\mathbf{p}^m\rangle_n = \mathbb{1}\{m = i\}, \forall m = 1, \dots, M. \tag{4}$$

If $n > M$ there are, generally speaking, infinitely many $\mathbf{a}$ satisfying (4). To choose one of them we adopt the minimax approach with the quadratic loss function. Then the risk of any estimate $\tilde{F}_i(A)$ for $F_i(A)$ is defined as

$$R_{\mathbf{F}}(\tilde{F}_i(A)) := \mathsf{E}_{\mathbf{F}}(\tilde{F}_i(A) - F_i(A))^2$$

and the assured (minimax) risk is

$$J(\tilde{F}_i(A)) := \sup_{\mathbf{F}} R_{\mathbf{F}}(\tilde{F}_i(A)), \tag{5}$$

where sup is taken over all possible sets of components' distributions. An estimate $\tilde{F}_i$ is minimax in some class of estimates if it minimizes the assured risk J over this class.

For weighted empirical distribution functions $\hat{F}_{\mathbf{a}}(A)$ with weights $\mathbf{a}$ satisfying the unbiasedness condition (4) we have

$$R_{\mathbf{F}}(\hat{F}_{\mathbf{a}}(A)) = \operatorname{Var} \hat{F}_{\mathbf{a}}(A) = \frac{1}{n^2}\sum_{j=1}^{n} (a_j)^2 \operatorname*{Var}_{\mathbf{F}}(\mathbb{1}\{\xi_j \in A\})$$

and

$$J(\hat{F}_{\mathbf{a}}(A)) = J(\mathbf{a}) = \frac{1}{4n}\langle \mathbf{a}, \mathbf{a}\rangle_n. \tag{6}$$

(Note that the sup in the definition of J is attained at any $\mathbf{F}$ with $F_m(A) = 1/2$ for all $m = 1, \dots, M$). So, to derive a minimax unbiased estimate for F_i we need to minimize $\langle \mathbf{a}, \mathbf{a}\rangle_n$ under the constrains (4).

Considering $\langle \cdot, \cdot \rangle_n$ as an inner product on $\mathbb{R}^n$ we see that the minimum is attained on $\mathbf{a}^i$ which is a linear combination of $\mathbf{p}^1$,…,$\mathbf{p}^M$ orthogonal to all $\mathbf{p}^m$ excepting $\mathbf{p}^i$. So

$$\mathbf{a}^i = \frac{\mathbf{p}^{i\perp}}{\langle \mathbf{p}^{i\perp}, \mathbf{p}^{i\perp}\rangle_n}, \tag{7}$$

where $\mathbf{p}^{i\perp}$ is an orthogonal complement of $\mathbf{p}^i$ to the linear subspace in $\mathbb{R}^n$ spanned by $\mathbf{p}^1, \dots, \mathbf{p}^{i-1}, \mathbf{p}^{i+1}, \dots, \mathbf{p}^M$. (This subspace will be denoted by $\mathcal{P}_{-i}$).

Of course, $\mathbf{a}^i$ exists if and only if $\mathbf{p}^i$ doesn't belong to $\mathcal{P}_{-i}$, i.e.

$$\langle \mathbf{p}^{i\perp}, \mathbf{p}^{i\perp} \rangle_n > 0. \tag{8}$$

A simple algebra shows that

$$\mathbf{a}^k = \mathbf{p}_{:n} \Gamma_n^+ \mathbf{e}_k, \tag{9}$$

where $\Gamma_n = (\langle \mathbf{p}^i \mathbf{p}^k \rangle_n)_{i,k=1}^M = \frac{1}{n} \mathbf{p}_{:n}^T \mathbf{p}_{:n}$, Γ_n^+ is the Moore-Penrose pseudoinverse for Γ_n, $\mathbf{e}_m = (\delta_{1m}, \dots, \delta_{Mm})^T$, $\delta_{ik} = \mathbb{1}\{i = k\}$. The assumption (8) is equivalent to

$$\Gamma_n \Gamma_n^+ \mathbf{e}_i = \mathbf{e}_i. \tag{10}$$

Under this assumption

$$J(\mathbf{a}^i) = \langle \mathbf{a}^i, \mathbf{a}^i \rangle_n = \frac{1}{4n} \mathbf{e}_i^T \Gamma_n^+ (\mathbf{p}_{:n})^T \mathbf{p}_{:n} \Gamma_n^+ \mathbf{e}_i = \frac{1}{4n} \mathbf{e}_i^T \Gamma_n^+ \mathbf{e}_i = \frac{1}{4n} \gamma_{ii}^+,$$

where γ_{ii}^+ is the (i,i)-th element of Γ_n^+. On the other hand, from (7) we get

$$J(\mathbf{a}^i) = \frac{1}{4n \langle \mathbf{p}^{i\perp}, \mathbf{p}^{i\perp} \rangle_n}$$

So we obtained the "minimax" weight array $\mathbf{a}^i$ for the estimation of F_i. The estimate $\hat{F}_i(A) = \hat{F}_{i:n}(A) = \hat{F}_{\mathbf{a}^i}(A)$ is minimax in the class of all unbiased weighted empirical measures. The following theorem states that $\hat{F}_i$ is minimax in the class of all unbiased estimates.

Theorem 1. *Assume that (10) holds and $\tilde{F}_i(A)$ is an unbiased estimate for F_i by the observations Ξ_n. Then*

$$J(\tilde{F}_i(A)) \geq J(\hat{F}_i(A)) = \frac{1}{4n} \gamma_{ii}^+.$$

(See Appendix for the proof).

Consistency of the minimax estimate can be demonstrated by the usual way applying the law of large numbers. A uniform consistency result analogous to the Glivenko-Cantelli theorem was also established for the weighted empirical measures and Vapnik-Červonenkis type inequality was obtained (see section 2.2. in Maiboroda and Sugakova, 2008).

4.2. **Moments estimation.** Basing on estimates for distributions one can construct estimates for many functionals from these distributions. E.g. if $g : \mathcal{X} \to \mathbb{R}$ is some nonrandom function then the functional moment

$$\bar{g}_i := \int g(x) F_i(dx) = \mathsf{E}\, g(\eta_i)$$

can be estimated by

$$\hat{g}_{i:n} := \int g(x) \hat{F}_{i:n}(dx) = \frac{1}{n} \sum_{j=1}^n a_{j:n}^i g(\xi_{j:n}). \tag{11}$$

This estimate is unbiased if $\bar{g}_m$ exist for $m = 1, \dots M$ and (10) holds.

Now let us formulate consistency conditions. In what follows we will not restrict ourselves by minimax weights only. So, let us define the weighted average

$$\hat{g}_{\mathbf{b}:n} := \frac{1}{n}\sum_{j=1}^{n} b_{j:n} g(\xi_{j:n}).$$

for any weight vector $\mathbf{b} = \{b_{j:n},\ j = 1, \dots, n, n = 1, 2 \dots\}$. Then $\hat{g}_{\mathbf{a}^i:n} = \hat{g}_{i:n}$.

Theorem 2. *Assume that for all* $k = 1, \dots, M$
(i) $\bar{g}_k$ *exist;*
(ii) $\langle \mathbf{b}\mathbf{p}^k \rangle$ *exist.*
(iii) $\sup_{j,n} |b_{j:n}| < \infty$.
Then $\hat{g}_{\mathbf{b}:n} \to \bar{g}_{\mathbf{b}} = \lim_{n\to\infty} \mathsf{E}\, \hat{g}_{\mathbf{b}:n} = \sum_{k=1}^{M} \langle \mathbf{b}\mathbf{p}^k \rangle \bar{g}_k$.

For the estimates $\hat{g}_{i:n}$ (ii) is equivalent to the existence of $\langle \mathbf{p}^i \mathbf{p}^k \rangle$ for all $i, k = 1, \dots, M$. Sometimes the mixing probabilities $\mathbf{p}$ are generated by some stochastic mechanism. Say, in Section 2, $p_j^1 = p(\tau_j)$ where τ_j are random variables. If p_j^i can be considered as i.i.d. for each i then $\langle \mathbf{p}^i \mathbf{p}^k \rangle = \mathsf{E}\, p_1^i p_1^k$ a.s. by the law of large numbers.

To assure (iii) one needs an asymptotic version of (8):

$$\liminf_{n\to\infty} \langle \mathbf{p}^{i\perp}, \mathbf{p}^{i\perp} \rangle_n > 0. \tag{12}$$

(Note that an asymptotic version of (10), say $\Gamma\Gamma^{+}\mathbf{e}_k = \mathbf{e}_k$, is not enough here since Γ^{+} is not a continuous function of Γ.) So Theorem 2 implies consistency of $\hat{g}_i$ as estimates of $\bar{g}_i$ if the assumption (12) holds.

If

$$\det \Gamma \neq 0, \tag{13}$$

then (12) holds for all i. So (13) is a natural condition of MVC model identifiability when all components' distributions (or their moments) are to be estimated. Assuming (12) we allow some uninteresting components to be unidentifiable and concentrate only on the i-th component analysis.

The proof of the theorem is based on the law of large numbers for series of random variables (theorem 3 from chapter 8 in Borovkov, 1998).

To formulate the asymptotic normality result we introduce some notations.

Assume that $\overline{(g)^2}_m := \int (g(x))^2 F_m(dx) < \infty$ for all $m = 1, \dots, M$. Denote

(14)

$$d_{j:n} := \operatorname{Var} g(\xi_{j:n}) = \mathsf{E}(g(\xi_{j:n}))^2 - (\mathsf{E}\, g(\xi_{j:n}))^2 = \sum_{m=1}^{M} \overline{(g)^2}_m p_{j:n}^m - \left(\sum_{m=1}^{M} \bar{g}_m p_{j:n}^m \right)^2.$$

Then

$$V_{:n}(\mathbf{b}) = V_{:n}(\mathbf{b}; \bar{g}_1, \dots, \bar{g}_M, \overline{(g)^2}_1, \dots, \overline{(g)^2}_M) := \operatorname{Var} \hat{g}_{\mathbf{b}:n} = \frac{1}{n} \langle (\mathbf{b})^2 \mathbf{d} \rangle_n = \frac{1}{n^2} \sum_{j=1}^{n} (b_{j:n})^2 d_{j:n}.$$

Theorem 3. *Assume that*
(i) $\overline{(g)^2}_m < \infty$ *for all* $m = 1, \dots, M$;
(ii) $\sup_{j,n} |b_{j:n}| < \infty$.
Then $(\hat{g}_{\mathbf{b}:n} - \mathsf{E}\, \hat{g}_{\mathbf{b}:n}) / \sqrt{V_{:n}(\mathbf{b})} \Rightarrow N(0, 1)$ *as* $n \to \infty$.

The proof is straightforward by applying the Lindenberg's version of the Central Limit Theorem (theorem 5 from chapter 8 in Borovkov, 1998).

Assumption (ii) holds for the minimax weight $\mathbf{b} = \mathbf{a}^m$ if (12) holds.

4.3. **Hypotheses testing.** Consistency and asymptotic normality of weighted averages $\hat{g}_{i:n}$ can be used to test hypotheses on the functional moments $\bar{g}_i$ for different mixture components. To simplify the presentation we consider only hypotheses of the form $H_0 : \bar{g}_m = \bar{g}_i$ for some fixed $g : \mathcal{X} \to \mathbb{R}$, m and i, against the general alternative $H_1 : \bar{g}_m \neq \bar{g}_i$. The proposed approach can be extended on hypotheses of much more general kind. (On testing some other homogeneity hypotheses see Maiboroda, 2000; Autin and Pouet, 2010).

It is natural to use a studentized version of the difference $D_{:n} = \hat{g}_{m:n} - \hat{g}_{i:n}$ as a test statistics. Observe that $D_{:n} = \hat{g}_{\mathbf{b}:n}$ with $\mathbf{b} = \mathbf{a}^m - \mathbf{a}^i$. Then applying Theorem 3 we see that under H_0

$$D_{:n}/\sqrt{V_{:n}(\mathbf{b})} \Rightarrow N(0,1).$$

So, to normalize $D_{:n}$ we need an estimate for $V_{:n}(\mathbf{b}) = \frac{1}{n}(V_{2:n}(\mathbf{b}) - V_{1:n}(\mathbf{b}))$, where

$$V_{2:n}(\mathbf{b}) := \frac{1}{n}\sum_{j=1}^{n} (b_{j:n})^2 \, \mathsf{E}(g(\xi_{j:n}))^2,$$

$$V_{1:n}(\mathbf{b}) := \frac{1}{n}\sum_{j=1}^{n} (b_{j:n})^2 \left(\mathsf{E}(g\xi_{j:n})\right)^2.$$

The value of $V_{2:n}$ can be approximated by its empirical counterpart

$$\hat{V}_{2:n}(\mathbf{b}) := \frac{1}{n}\sum_{j=1}^{n} (b_{j:n})^2 (g(\xi_{j:n}))^2.$$

The estimation of $V_{1:n}$ is less evident. Observe that $V_{1:n}(\mathbf{b}) = \bar{\mathbf{g}}^T \Gamma_{\mathbf{b}:n} \bar{\mathbf{g}}$, where $\bar{\mathbf{g}} := (\bar{g}_1, \dots, \bar{g}_M)^T$ and

$$\Gamma_{\mathbf{b}:n} := \frac{1}{n}\sum_{j=1}^{n} (b_{j:n})^2 \mathbf{p}_{j:n} \mathbf{p}_{j:n}^T.$$

To estimate $\bar{g}_k$, $k = 1, \dots, M$ we use $\hat{g}_{k:n}$ defined by (11) with the weights $\mathbf{a}^k$ from (9). Note that the estimates $\hat{g}_{k:n}$ can be inconsistent if (12) doesn't hold with $i = k$. But we will show that $\hat{V}_{1:n} = \hat{\mathbf{g}}^T \Gamma_{\mathbf{b}:n} \hat{\mathbf{g}}$ is a good approximation to $V_{1:n}(\mathbf{b})$ even in this case.

So our approximation to $V_{:n}(\mathbf{b})$ is $\hat{V}_{:n} = \frac{1}{n}(\hat{V}_{2:n} - \hat{V}_{1:n})$ and the studentized statistics for testing H_0 is $T_{:n} = D_{:n}/\sqrt{\hat{V}_{:n}}$.

Theorem 4. *Assume that H_0 holds and*
(i) $\overline{(g)^2}_k < \infty$ for all $k = 1, \dots M$,
(ii) $\langle \mathbf{p}^{k_1} \mathbf{p}^{k_2} \rangle$ and $\langle \mathbf{p}^{k_1} \mathbf{p}^{k_2} \mathbf{b} \rangle$ exist for all $k_1, k_2 = 1, \dots, M$;
(iii) assumption (12) holds for $k = m$ and $k = i$,
(iv) $V_{:\infty} = \lim_{n\to\infty} nV_{:n}(\mathbf{b}) > 0$.
Then $T_{:n} \Rightarrow N(0,1)$.

(See Appendix for the proof).

So the test $\mathcal{T}_{:n}$ which rejects H_0 iff $|T_{:n}| > \Phi^{-1}(1 - \alpha/2)$ has the asymptotic significance level α. The p-level of this test is $p^* = 2(1 - \Phi(|T_{:n}|))$. Here $\Phi^{-1}(x)$ is the probit transform, i.e. the function inverse to the standard normal CDF $\Phi(x)$.

Significance of $\mathcal{T}_{:n}$

F_2	n				
	100	250	500	750	1000
$N(0,1)$	0.0597	0.0518	0.0543	0.0504	0.0506
$N(0,4)$	0.0576	0.0528	0.053	0.0497	0.0491
T_3	0.0533	0.0476	0.0497	0.0491	0.0501

Variance of $T_{:n}$

F_2	n				
	100	250	500	750	1000
$N(0,1)$	1.081	1.011	1.010	0.999	0.992
$N(0,4)$	1.053	1.033	1.028	1.002	0.981
T_3	1.032	1.001	0.989	0.999	1.000

TABLE 1. Results of simulation

4.4. **Density estimation.** Weighting with minimax weights can be also used to obtain estimates of components characteristics different from functional moments. If we assume that $\mathcal{X} = \mathbb{R}$ and all the components distributions are absolutely continuous then the density f_m of the distribution F_m may be estimated by the weighted kernel density estimate

$$\hat{f}_{m:n}(x) = \frac{1}{hn} \sum_{j=1}^{n} a_{j:n}^{m} K\left(\frac{x - \xi_{j:n}}{h}\right),$$

where K is a kernel (i.e. a probability density) and $h > 0$ is a bandwidth. Asymptotic theory of such estimates and approaches to optimal choice of the kernel and bandwidth are rather similar to these for the usual kernel density estimate based on i.i.d. observations. It was developed in Sugakova (1999). On projective density estimates with wavelet basis see Pokhyl'ko (2005).

5. Numerical examples

5.1. **Simulations results.** We have performed a small simulation study to see how far is the test statistics behavior on samples of small and moderate sizes from the asymptotic results obtained in Section 4.3. We considered two components mixtures with the mixing probabilities $p_{j:n}^1 = j/n$ and $p_{j:n}^2 = 1 - p_{j:n}^1$. The test $\mathcal{T}_{:n}$ was applied to test the hypotheses of means homogeneity, i.e. $H_0 : \mathbb{E}\eta_1 = \mathbb{E}\eta_2$. In all considered examples the first component of the mixture η_1 was standard normal. The second component η_2 was taken $N(0,1)$ in the first example, so the observations were i.i.d. here. In the second example we took $N(0,4)$ as the second component to see how the difference of variances affects the test of means homogeneity. In the third example we investigated the influence of heavy tails taking the Student-T distribution with 3 degrees of freedom for the second component.

Results of simulations are presented in Table 1. Here the empirical significance levels of the test $\mathcal{T}_{:n}$ with nominal significance level 0.05 and the empirical variances of $T_{:n}$ are given, obtained over 10 000 simulated samples.

These results demonstrate satisfactory performance of the test for sample sizes $n \geq 750$.

5.2. **Real data example.** In Hedenfalk et al. (2001) a study of more than 3000 of genes was performed for 22 tissue specimens taken from breast cancer tumors:

— 7 specimens from BCRA1-positive tumors,

— 8 specimens from BCRA2-positive tumors,

— 7 specimens from "sporadic" tumors.

Here BCRA1 and BCRA2 are specific mutations which cause malignant transformation. The sporadic tumors are not connected to any specified mutation but were sorted out by their clinical features.

As the result of DNA microarray analysis the set of genes expression levels x_{ik}^j is obtained, where

— j is the number of a considered gene, $j = 1, \dots, 3170$;

— i denotes the type of a tissue ($i = 1$ for BCRA1-positive specimens, $i = 2$ for BCRA2-positive ones, $i = 3$ for sporadic tumors);

— k is the number of a specimen in the sample.

One of the main goals of the researchers was to identify genes with different mean expression levels in BCRA1 and BCRA2 tumors.

The standard way to decide if the j-th gene is/isn't equally expressed in two types of tumors is to apply the two-sample Student-T test (or Fisher-F test) to compare the samples $\{x_{1k}^j,\ k = 1, \dots, 7\}$ and $\{x_{2k}^j,\ k = 1, \dots, 8\}$. As a result, one obtains the p-level π_j of the test. Then

— π_j is uniformly [0,1] distributed if the hypothesis H_0^j holds, where H_0^j: *the mean expression level of the j-th gene is the same for BRCA1 and BRCA2 positive tumors.*

— $\pi_j \in [0, 1]$ is nearly zero with high probability if the alternative H_1^j holds, where H_1^j: *mean expression levels of the j-th gene are different.*

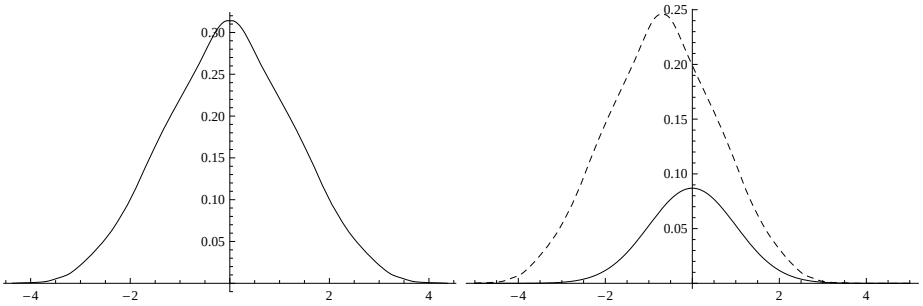
Then fixing the significance level α, one accepts H_1^j and attributes a gene j as an interesting (differently expressed) one if $\pi_j < \alpha$. If $\pi_j \geq \alpha$ the gene is considered as an uninteresting (equally expressed) one, i.e. H_0^j is accepted.

In this procedure one controls the first type error (α) only. I.e. we may say that the proportion of uninteresting genes erroneously considered as interesting ones is $\leq \alpha$. But how many interesting genes will be discarded? To answer this question we need to know the distribution of π_j under the alternative. Note that biologists prefer to work not with the p-levels π_j, but with Z-scores $Z_j = \Phi^{-1}(\pi_j)$. We will see now how to estimate the distribution of Z_j.

So the distribution of Z_j for uninteresting genes is $N(0, 1)$. Following Bordes et al. (2006) we assumed that the distribution of Z_j for interesting genes is symmetric around its (unknown) median a. Then the distribution of observed data Z_j is a mixture of two components with the density

$$f^Z(x) = qf(x - a) + (1 - q)\varphi(x),$$

where q is the mixing probability, f is the even density of $Z_j - a$ for interesting genes, φ is the standard normal density. Estimation of parameters in such models is discussed in Bordes et al. (2006), Maiboroda and Sugakova (2011). In Maiboroda and Sugakova (2011a) we estimated a, p and f by the Hedenfalk data and obtained the estimates $\hat{a} = -0.69$, $\hat{q} = 0.78$. The graph of the density estimate $\hat{f}$ is presented in Fig. 1(a). These results are rather discouraging: we obtained that the sample contains nearly 78% of interesting genes (which change their expression in different tumors) and 22% of uninteresting ones. If one applied the



(a) (b)

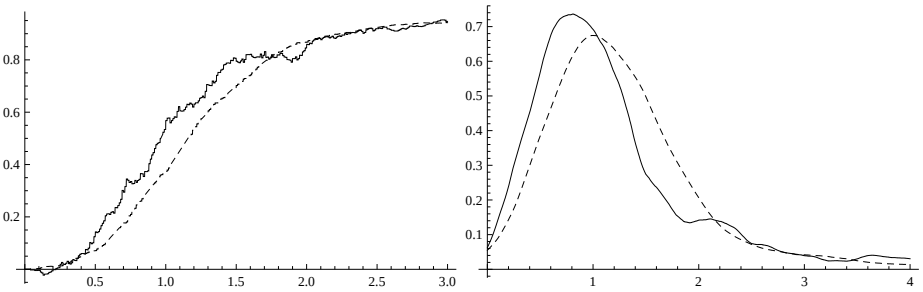
FIGURE 1. Densities for Z-values for Hedenfalk data (a) Symmetric density f, (b) Densities of components multiplied by mixing probabilities (dashed line for the interesting component, solid line for the uninteresting one).

procedure of interesting genes selection as described above, with, say, $\alpha = 0.05$, then $\int_{\Phi^{-1}(\alpha)} f(x-a)dx \approx 76\%$ of interesting genes would not be recognized adequately. If one applied an empirical Bayes classifier based on the estimated densities of the components $\hat{f}_1^Z(x) = \hat{f}(x-\hat{a})$ and $f_2^Z(x) = \varphi(x)$ then all the genes would be classified as interesting since $\hat{q}\hat{f}_1^Z(x) > (1-\hat{q})f_2^Z(x)$ for all x, see Fig. 1 (b).

Therefore we propose to use the approach described in Section 2 to analyze differences in expression of interesting and uninteresting genes in sporadic tumors. The results for the first specimen of this kind of tumors (denoted by Sp1) are presented in Fig. 2. Here the distribution of interesting genes seems shifted to the right with respect to the distribution of uninteresting ones. Is this shift enough to cause a significant change in means? Applying the test $\mathcal{T}_{:n}$ from Section 4.3 with $g(x) = x$ we obtained $T_{:n} = 0.776135$ and p-value 0.437669. So the difference of means is nonsignificant.

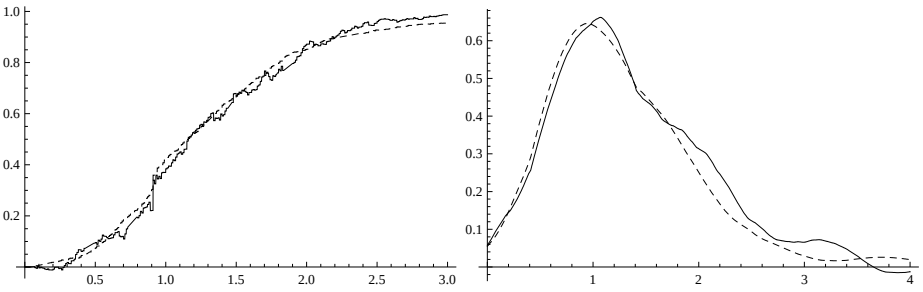
Does it mean that the difference of distributions observed at Fig 2 is absolutely insignificant? Surely no. In fact, we see that the estimated probabilities for the expression levels to be less than 1 are quite different. Applying the test $\mathcal{T}_{:n}$ with $g(x) = \mathbb{1}\{x < 1\}$ to test significance of this difference one obtains $T_{:n} = 2.19461$ and p-level 0.0281916. With the standard significance level $\alpha = 0.05$ we conclude that the distributions of expression levels are significantly different for interesting and uninteresting genes in Sp1.

In contrast to this, analogous estimates for another specimen of sporadic tumor (marked as Sp7) don't reveal any significant difference between distributions of interesting and uninteresting genes expression levels (see Fig. 3). The test $\mathcal{T}_{:n}$ for the means homogeneity yields $T_{:n} = 0.915729$ and p-value 0.665829. We conclude



(a) (b)

FIGURE 2. Distribution of expression levels in Sp1 (a) Cumulative distribution functions, (b) Densities (dashed line for the interesting genes, solid line for the uninteresting ones).



(a) (b)

FIGURE 3. Distribution of expression levels in Sp7 (a) Cumulative distribution functions, (b) Densities (dashed line for the interesting genes, solid line for the uninteresting ones).

that the genes of both kinds demonstrate the same distribution of expression levels in this specimen.

The obtained results are not surprising. The class of sporadic tumors was defined not by some genetical features (as the BCRA1 and BCRA2 tumors were) but by clinical observations. So different genetical causes of these tumors are possible.

The proposed technique can make these causes more comprehensible. Of course, the considered two examples are not enough for general conclusions, but we hope that genetical profiles similar to the represented above may be used as diagnostic tools for individual patients.

6. Conclusion

We discussed techniques of estimation and hypotheses testing based on the model of finite mixture with varying concentrations (mixing probabilities). It is shown that this model allows making substantial conclusions on some mixture components even if some other components distributions are unidentifiable. This technique may be applied to the analysis of DNA microarrays data.

7. Appendix

Proof of Theorem 1. To simplify notations let $i = 1$. We will consider a parametric submodel of (2) and apply the Cramér-Rao lower bound to get the statement of the theorem. Take a set $\mathbf{u}^1, \dots, \mathbf{u}^K$ $(K \le M)$ of vectors in $\mathbb{R}^n$, such that $\mathbf{p}^1 = \mathbf{u}^1$, $\mathbf{p}^m = \sum_{k=2}^K c_{mk}\mathbf{u}_k$ for $m = 2, \dots, M$, $\langle \mathbf{u}^k, \mathbf{u}^l \rangle_n = \delta_{kl}$, $k, l = 2, \dots, K$. Here c_{mk}, $m = 2, \dots, M$, $k = 2, \dots, K$ are some constants.

Let x_1 and x_2 be any two different points in $\mathcal{X}$. Consider a set of distributions F on $\mathcal{X}$ concentrated on $\{x_1, x_2\}$ and restrict ourselves by F_i from this set, i.e.

$$F_i(A) = f_i \mathbb{1}\{x_1 \in A\} + (1 - f_i)\mathbb{1}\{x_2 \in A\}, \tag{15}$$

where $0 \le f_i \le 1$ are some parameters. Then by (2)

$$\Psi_j(A) = \sum_{m=1}^M p_j^m F_m(A) = u_j^1 F_1(A) + \sum_{m=2}^M \sum_{k=2}^K c_{mk} u_j^k F_m(A)$$

and

$$\Psi_j(A) = (\sum_{k=1}^K u_j^k t_k)\mathbb{1}\{x_1 \in A\} + (1 - \sum_{k=1}^K u_j^k t_k)\mathbb{1}\{x_2 \in A\}, \tag{16}$$

where $t_1 = f_1$, $t_k = \sum_{m=2}^M c_{mk} f_m$. We consider (16) as a parametric model with the unknown parameter $\vartheta = (t_1, \dots, t_K)$. Let the true values of the parameters be $t_1 = 1/2$, $t_k = \sum_{m=2}^M c_{mk}/2$. (These t_k correspond to $f_k = 1/2$). The entries of the information matrix $I^j = (I_{kl}^j)_{k,l=1}^K$ for the information on the parameter ϑ contained in the observation ξ_j can be evaluated as

$$I_{kl}^j = \frac{u_j^k u_j^l}{\sum_{i=1}^K u_j^i t_i} + \frac{u_j^k u_j^l}{1 - \sum_{i=1}^K u_j^i t_i} = 4 u_j^k u_j^l$$

since $\sum_{i=1}^K u_j^i t_i = \sum_{m=1}^M p_m f_m = \frac{1}{2}$.

So the information matrix for the full sample Ξ_n is

$$I = \sum_{j=1}^n I^j = 4n \begin{pmatrix} \langle \mathbf{p}^1, \mathbf{p}^1 \rangle_n & \langle \mathbf{p}^1, \mathbf{u}^2 \rangle_n & \langle \mathbf{p}^1, \mathbf{u}^3 \rangle_n & \dots & \langle \mathbf{p}^1, \mathbf{u}^K \rangle_n \\ \langle \mathbf{p}^1, \mathbf{u}^2 \rangle_n & 1 & 0 & \dots & 0 \\ \langle \mathbf{p}^1, \mathbf{u}^3 \rangle_n & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{p}^1, \mathbf{u}^K \rangle_n & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Then $\det I = 4n(\langle \mathbf{p}^1, \mathbf{p}^1\rangle_n - \sum_{k=2}^K \langle \mathbf{p}^1, \mathbf{u}^k\rangle_n^2) = 4n\langle \mathbf{p}^{1\perp}, \mathbf{p}^{1\perp}\rangle_n$ and the (1,1)-th element of I^{-1} is $1/\det I = \gamma_{11}^+/(4n)$. So, by the Cramér-Rao inequality, for any unbiased estimate $\tilde{t}_1$ of t_1, $\mathsf{E}(\tilde{t}_1 - t_1)^2 \geq \gamma_{11}^+/(4n)$. But in our parametric model $t_1 = F_1(A)$ if $x_1 \in A$, $x_2 \notin A$ and $\tilde{F}_1(A)$ is an unbiased estimate of t_1. Then $J(\tilde{F}_1) \geq \mathsf{E}_{\mathbf{F}}(\tilde{F}_1(A) - F(A))^2 \geq \gamma_{11}^+/(4n)$ where $\mathbf{F}$ is a set of F_i defined by (15) with $f_i = 1/2$.

□

Proof of Theorem 4. By (ii) Γ and $\Gamma_{\mathbf{b}} = \lim_{n\to\infty} \Gamma_{\mathbf{b}:n}$ exist and by (iii) $\sup_{j,n} |a^k_{j:n}| < \infty$ for $k = i$ and $k = m$. Then $V_{:\infty} := \langle (\mathbf{b})^2 \mathbf{d}\rangle$ exists and by Theorem 3,

$$D_{:n}/\sqrt{nV_{:\infty}} \Rightarrow N(0,1). \tag{17}$$

To complete the proof we need to show that $\hat{V}_{:n} \to V_{:\infty}$ in probability. Note that $V_{:\infty} = V_{2:\infty} - V_{1:\infty}$, where

$$V_{2:\infty} = \sum_{k=1}^M \langle (\mathbf{b})^2 \mathbf{p}^k\rangle \overline{(g)^2}_k,\ V_{1:\infty} = \bar{\mathbf{g}}^T \Gamma_{\mathbf{b}} \bar{\mathbf{g}}.$$

By Theorem 2, $\hat{V}_{2:n} \to V_{2:\infty}$ in probability as $n \to \infty$. Similarly,

$$\hat{\mathbf{g}}_{:n} - \mathsf{E}\,\hat{\mathbf{g}}_{:n} \to 0, \text{ in probability as } n \to \infty. \tag{18}$$

But $\mathsf{E}\,\hat{\mathbf{g}}_{:n} = \frac{1}{n}\Gamma^+_{:n}\mathbf{p}^T_{:n}\mathbf{p}_{:n}\bar{\mathbf{g}} = \Gamma^+_{:n}\Gamma_{:n}\bar{\mathbf{g}}$ may be not converging to $\bar{\mathbf{g}}$ if Γ is singular.

On the other hand, the operator $\boldsymbol{\pi}_{:n} = \Gamma^+_{:n}\Gamma_{:n}$ is a projector of $\mathbb{R}^n$ onto the orthogonal complement to $\mathrm{Ker}(\Gamma_{:n})$ (the null space of $\Gamma_{:n}$). But $\mathrm{Ker}(\Gamma_{:n}) \subseteq \mathrm{Ker}(\Gamma_{\mathbf{b}:n})$. Really, since $\Gamma_{:n}$ is a symmetric nonnegative matrix, a vector $\mathbf{c} \in \mathrm{Ker}(\Gamma_{:n})$ iff $\mathbf{c}^T\Gamma_{:n}\mathbf{c} = \|\mathbf{p}_{:n}\mathbf{c}\|^2_{:n} = 0$. Then $|\mathbf{c}^T\Gamma_{\mathbf{b}:n}\mathbf{c}| = |\mathbf{c}^T\mathbf{p}^T_{:n}\mathbf{b}_{:n}\mathbf{p}_{:n}\mathbf{c}| \leq \sup_{j,n}|b_{j:n}| \cdot \|\mathbf{p}_{:n}\mathbf{c}\|^2_{:n} = 0$, i.e. $\mathbf{c} \in \mathrm{Ker}(\Gamma_{\mathbf{b}:n})$.

So $\boldsymbol{\pi}_{:n}\Gamma_{\mathbf{b}:n} = \Gamma_{\mathbf{b}:n}\boldsymbol{\pi}_{:n} = \Gamma_{\mathbf{b}:n}$ and $(\mathsf{E}\,\hat{\mathbf{g}}_{:n})^T\Gamma_{\mathbf{b}:n}\,\mathsf{E}\,\hat{\mathbf{g}}_{:n} = V_{1:n} \to V_{1:\infty}$ as $n \to \infty$. Taking (18) into account we get $\hat{V}_{1:n} \to V_{1:\infty}$ and $\hat{V}_{:n} \to V_{:\infty}$ in probability as $n \to \infty$. This with (17) implies the statement of the theorem.

□

References

[1] Autin F., Pouet Ch. (2010) Test on the components of mixture densities. *eprint arXiv:0912.0786.*

[2] Bordes L., Mottelet S., Vandekerkhove P. (2006). Semiparametric estimation of a two-component mixture model.*Ann. Statist.* **34**, 1204-1232.

[3] Bordes L., Delmas C., Vandekerkhove P. (2006a). Semiparametric Estimation of a two-component Mixture model where one component is known.*Scand. J. Statist.* **33**, 733-752.

[4] Borovkov A.A. (1998) *Probability Theory.* Gordon and Breach Science Publishers, Amsterdam.

[5] Hall P., Zhou X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann.Statist.*, **31**, 201-224.

[6] Holzmann, H., Munk, A. & Gneiting T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Statist.* **33**, 753-763.

[7] Hunter D.R., Wang S., Hettmansperger T.R. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35**, 224-251.

[8] Lodatko A., Maiboroda R. (2006) Estimation of the probability density by observations with an admixture.*Theory of Probability and Mathematical Statistics* **73**, 99-108.

[9] Maiboroda R.E. (1993) Projection estimators of varying concentrations of mixtures. *Theory of Probability and Mathematical Statistics* **64**, 71-75.

[10] Maiboroda R.E. (2000) A homogeneity criterion for mixtures with varying concentrations. *Ukrainian Math. J.* **52**, 1256-1263.

[11] Maiboroda R.E. (2002) Least square estimates for parameters of concentrations of varying mixtures. I. The consistency. *Theory of Probability and Mathematical Statistics* **64**, 105-115.

[12] Maiboroda, R. (2008). Estimation of locations and mixing probabilities by observations from two-component mixture of symmetric distributions. *Teorija Imovirnosti ta Matematychna Statystyka* **78**, 133-141 (in Ukrainian).

[13] Maiboroda, R., Sugakova O. (2008) Estimation and classification by observations from mixtures. Kyiv University Publishers, (in Ukrainian).

[14] Maiboroda R., Sugakova O. (2011) Generalized Estimating Equations for Symmetric Distributions Observed with Admixture, *Communications in Statistics - Theory and Methods*, 40: 1, 96 116.

[15] Maiboroda R., Sugakova O. (2011a) Nonparametric density estimation for symmetric distributions by contaminated data, *Metrika* to appear.

[16] McLachlan, G.J., Peel, D. (2000). *Finite Mixture Models.* Wiley, New York.

[17] Pokhyl'ko D. (2005) Wavelet estimators of a density constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics* **70**, 135-145.

[18] Shao, J. (1998). *Mathematical statistics.* Springer-Verlag, New York.

[19] Sugakova O.V. (1999) Asymptotics of a kernel estimate for distribution density constructed from observations of a mixture with varying concentration. *Theory of Probability and Mathematical Statistics* **59**, 161-171.

[20] Tanaka, K. (2009) Strong consistency of the minimum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scalar parameter. *Scand. J. Statist.*, **26**, 171-184.

[21] Titterington, D.M., Smith, A.F., Makov, O.E. (1985). *Analysis of Finite Mixture Distributions.* Wiley, New York.

Kyiv National University, Ukraine
E-mail address: mre@univ.kiev.ua