УДК 519.21

О.В. Доронін[1], аспірант

Тест для перевірки рівності розподілів у моделі суміші зі змінними концентраціями

[1]Київський національний університет імені Тараса Шевченка, 83000, м. Київ, пр-т. Глушкова, 4е,
e-mail: al_doronin@ukr.net

O.V. Doronin[1], PhD Student

Test for checking the equality of distributions in mixture model with varying concentrations

[1]Taras Shevchenko National University of Kyiv, 83000, Kyiv, Glushkova st., 4e,
e-mail: al_doronin@ukr.net

У даній роботі розглядається модель суміші ймовірнісних розподілів, у якій концентрації вважаються відомими, і змінюються від спостереження до спостереження. Для даної моделі ставиться задача перевірки гіпотези про рівність розподілів компонентів між собою, або деякому наперед заданому розподілу. В якості розв'язку пропонується техніка, аналогічна класичному тесту хі-квадрат, де рівність розподілів перевіряється по групованим даним. Дана техніка є модифікацією більш загального підходу до перевірки гіпотез про функціональні моменти у моделі суміші, який був розглянутий у попередній роботі автора. Для конкретно даної модифікації конкретизовано умови її застосування на практиці. На роль тестової статистики обирається вектор, елементами якого є значення навантаженої емпіричної міри на деяких підмножинах простору спостережень. Тест полягає у дослідженні відстані у сенсі Махаланобіса між значенням тестової статистики та нулем. Дана відстань порівнюється із пороговим рівнем, який виводиться із асимптотичних властивостей статистики. Наведений підхід легко може бути узагальнений для перевірки більш широкого класу гіпотез про розподіли компонентів суміші. Якість роботи тесту перевіряється за допомогою імітаційного моделювання.

Ключові слова: модель суміші, перевірка гіпотез, тест хі-квадрат.

In this paper we consider the model of mixture of probability distributions. The concentrations of components (mixing probabilities) are known, and they vary from observation to observation. For the given model, we formulate a problem of testing hypothesis about the equality of components' distributions within each other, or to some distribution given in advance. As a solution we propose a technique analogous to the classical chi-square test, where distributions' equality is checked by the grouped data. This technique is a modification of more general approach to testing hypotheses about functional moments in the model of mixture, which was developed in the author's previous work. For this concrete modification, the conditions of it's application were concretized. As a test statistic, we choose a vector, which elements are the values of the weighted empirical measure on some subsets of the observations' space. To perform the test, we compare the Mahalanobis-type distance between the value of the test statistic and zero. This distance is compared with the threshold value, which we derive from the asymptotic properties of statistic. Introduced approach can easily be generalyzed for checking of much more general class of hypotheses about the mixture components' distributions. Performance of the test is analyzed by simulation.

Key Words: mixture model, hypothesis testing, chi-square test.

Communicated by Prof. Kozachenko Yu.V.

Introduction

We consider a sample consisting of N subjects $O_{1;N}, ..., O_{N;N}$. Each subject $O_{j;N}$ belongs to one of the M subpopulations $\mathcal{P}_1, ..., \mathcal{P}_M$ (mixture's components) of the general population. But we do not know to which subpopulation exactly. We denote the number of this subpopulation as $\text{ind}(O_{j;N}) \in \{1, ..., M\}$. Next we assume that we know only the probabilities that $\text{ind}(O_{j;N})$ takes certain value (component's concentrations):

$$\begin{aligned} p^m_{j;N} &:= \mathbf{P}[\text{ind}(O_{j;N}) = m] \\ &= \mathbf{P}[O_{j;N} \in \mathcal{P}_m]. \end{aligned} \tag{1}$$

We observe some characteristics $\xi_{j;N} := \xi(O_{j;N})$ for each subject $O_{j;N}$. Assume next that the values of ξ lie in some measurable space $\mathfrak{X}$ with $\mathfrak{F}$ as it's σ-algebra. Denote as $F_m(A) := \mathbf{P}[\xi(O) \in A | O \in \mathcal{P}_m]$ the conditional distribution of $\xi(O)$ assuming that the subject O is taken from the m-th mixture's subpopulation. Then the distribution of $\xi_{j;N}$ is expressed as

$$\mathbf{P}[\xi_{j;N} \in A] = \sum_{m=1}^{M} p_{j;N}^{m} F_m(A), \ A \in \mathfrak{F}. \qquad (2)$$

The mixture model with varying concentrations may appear in the analysis of medical and biological data (see [11]), during sociological and politological researches (see [12]), and in scope of economical, psychological and other issues (see [5]).

In this paper we consider testing hypotheses about equality of distributions for different mixture's components. As a solution of this problem, it is proposed a test analogous to the classical chi-square test. This test checks the equality of distributions by the grouped data. This paper is the sequel of [3], where more general scheme of testing hypotheses on functional moments was considered. The test of checking the single-dimentional functional moments for two mixture's components was developed in [11]. The hypothesis of the homogeneity of two different samples is considered in [6].

Another problems for the mixture's model are considered in [1, 2, 4, 14]. Note that the mixture's model with varying concentration is a modification of the classical mixture's model. In the last one, probabilities $p_{j;N}^{m}$ are the same for all $j = 1, ..., N$. The monographs [13, 15] are devoted to the classical mixture's model. Nonparametric technique is applied to this model in [7, 8, 10].

In this paper, the formal problem statement is placed in section 1. Test statistics are introduced in section 2. We investigate the asymptotic properties of test statistics more thoroughly in section 3. The test itself is constructed in section 4. The results of simulation study is placed in section 5.

1 Problem statement

Our goal is to develop a test for checking the hypotesis about the equality of some component's distribution to the distribution of some another component's distribution. But in fact our test will check the equality of the grouped data distributions. I.e. we check the equality of distributions F_a i F_b on some sets $A_1, ..., A_Q$ from $\mathfrak{F}$:

$$F_a(A_q) = F_b(A_q), \quad q = 1, ..., Q. \qquad (3)$$

Without loss of generality we check the hypothesis about only first mixture's components. Indeed, the mixture's components can be rearranged without change of (2). The first type of hypotheses we are interested in is the hypothesis of the equality of F_1 to some distribution F_0 given in advance:

$$H_0 : \ F_1(A_q) = F_0(A_q), \ q = 1, ..., Q. \qquad (4)$$

The second type of hypotheses is about the equality of distributions of the first two mixture's components:

$$H_0 : \ F_1(A_q) = F_2(A_q), \ q = 1, ..., Q. \qquad (5)$$

In what follows we build the test to check both types of hypotheses, (4) and (5). This test can be considered as the analogue of chi-square test for the model of mixture with varying concentrations.

2 Test statistics

Hereinafter we denote zero vector from $\mathbb{R}^k$ as $\mathbb{O}_k$. Unit k-by-k matrix is denoted as $\mathbb{I}_{k\times k}$, and zero k-by-m matrix as $\mathbb{O}_{k\times m}$. diag$[v]$ will denote the diagonal matrix, the diagonal elements of which are the elements of some vector v. For real-valued m-by-k matrix A we will write $A \in \mathbb{R}^{m\times k}$. We denote the convergence in probability as $\xrightarrow{P}$. And the convergence by distribution as $\xrightarrow{d}$. Moreover, we introduce independent random variables η_m with distributions F_m, $m = 1, ..., M$. For arbitraty sets of values $a_{j;N}$ and $b_{j;N}$, $j = 1, ..., N$ we define the averaging operator $\langle\cdot\rangle_N$:

$$\langle a_{;N} \rangle_N := \frac{1}{N} \sum_{j=1}^{N} a_{j;N}. \qquad (6)$$

The operations of summation, multiplication, etc. under the operator $\langle\cdot\rangle_N$ we understand elementwise:

$$\begin{aligned} \langle a_{;N} + b_{;N} \rangle_N &= \frac{1}{N} \sum_{j=1}^{N} (a_{j;N} + b_{j;N}), \\ \langle a_{;N} b_{;N} \rangle_N &= \frac{1}{N} \sum_{j=1}^{N} a_{j;N} b_{j;N}. \end{aligned} \qquad (7)$$

Corresponding limit values (if they exist) we denote through the operator $\langle\rangle$ without subscript N:

$$\langle a * b\rangle := \lim_{N\to\infty}\langle a_{;N} * b_{;N}\rangle_N. \tag{8}$$

Let us assume that $A_1, ..., A_Q$ are some measurable sets from $\mathfrak{F}$. We denote the values of mixture's component's distributions on these sets as

$$f_q^m := F_m(A_q),\ q = 1, ..., Q,\ m = 1, ..., M, \tag{9a}$$

$$f_{r,s}^m := F_m(A_r \cap A_s),\ r, s = 1, ..., Q,\ m = 1, ..., M. \tag{9b}$$

Then we can reformulate hypotheses (4) and (5) as $f^1 = f^0$ and $f^1 = f^2$ respectively, where $f^0 := \big(F_0(A_q)\big)_{q=1,...,Q}$ is the vector of the values of given in advance distribution on the sets A_q.

We will use the weighted empirical measures $\hat{F}_{m;N}$ to estimate f_q^m i $f_{r,s}^m$:

$$\hat{F}_{m;N}(A) := \frac{1}{N}\sum_{j=1}^{N} a_{j;N}^m \mathbb{I}_{\xi_{j;N}\in A},\ A \in \mathfrak{F}, \tag{10}$$

where $a_{j;N}^m$ is the set of some weight coefficients. We define $a_{j;N}^m$ as the minimax weight coefficients (e.g., see [3], [5]):

$$a_{\cdot;N}^m := p_{\cdot;N}\Gamma_N^{-1} e_m, \tag{11}$$

where $p_{\cdot;N} := (p_{j;N}^m)_{j=1,...,N, m=1,...,M} \in \mathbb{R}^{N\times M}$ is the matrix of concentrations, Γ_N^{-1} is the inversed matrix to the Gramm's matrix of concentrations

$$\Gamma_N := \big(\langle p_{;N}^k p_{;N}^l\rangle_N\big)_{k,l=1,...,M} \in \mathbb{R}^{M\times M}, \tag{12}$$

and $e_m := \big(\mathbb{I}_{\{k=m\}}\big)_{k=1,...,M} \in \mathbb{R}^M$. Thus, the values of f_q^m and $f_{r,s}^m$ can be estimated as

$$\hat{f}_q^m := \hat{F}_{m;N}(A_q),\ \hat{f}_{r,s}^m := \hat{F}_{m;N}(A_r \cap A_s). \tag{13}$$

If the space of observations is real-valued ($\mathfrak{X} = \mathbb{R}$) with Borel σ-algebra $\mathfrak{F} = \mathfrak{B}(\mathbb{R})$ we can take the improved weight coefficients $\tilde{a}_{j;N}^m$ (see [3], [5], [9]). We can define them as the coefficients that correspond to improved empirical distribution function

$$\begin{aligned}\tilde{F}_{m;N}(x) &:= \min\{1, \sup_{y<x}\hat{F}_{m;N}(y)\}\\ &= \frac{1}{N}\sum_{j=1}^{N}\tilde{a}_{j;N}^m \mathbb{I}_{\xi_{j;N}\le x}.\end{aligned} \tag{14}$$

Theorem 1 shows consistency and asymptotic normality of estimators $\hat{F}_{m_0;N}$ and $\tilde{F}_{m_0;N}$. To formulate the theorem we will denote the limit value of Γ_N as

$$\Gamma := \lim_{N\to\infty}\Gamma_N = \big(\langle p^k p^l\rangle\big)_{k,l=1,...,M}. \tag{15}$$

Theorem 1. *Let B_k be some measurable sets from $\mathfrak{F}$, m_k be the indices from $\{1, ..., M\}$, $k = 1, ..., K$, and fulfills the following.*
(i) Limit matrix Γ exists, and $\det\Gamma \neq 0$.
(ii) Limit values $\langle a^{m_k} a^{m_l} p^r p^s\rangle$ exist, $r, s = 1, ..., M$, $k, l = 1, ..., K$.
Then 1. $\sqrt{N}\big(\hat{F}_{m_k;N}(B_k) - F_{m_k}(B_k)\big)_{k=1,...,K} \xrightarrow{d} \zeta \simeq \mathcal{N}(\mathbb{O}_K, \Sigma)$ as $N \to \infty$, where $\hat{F}_{m_k;N}$ is the empirical measure from (10)*, and Σ is the dispersion matrix with elements defined as*

$$\begin{aligned}(\Sigma)_{k,l} &= \sum_{m=1}^{M}\langle a^{m_k} a^{m_l} p^m\rangle F_m(A_k \cap A_l)\\ &- \sum_{r,s=1}^{M}\langle a^{m_k} a^{m_l} p^r p^s\rangle F_r(A_k) F_s(A_l).\end{aligned} \tag{16}$$

2. If moreover the following conditions are fulfilled,
(iii) Distribution functions F_m are continuous on $\mathbb{R}$, $m = 1, ..., M$.
(iv) $\operatorname{supp} F_m \subseteq \operatorname{supp} F_{m_k}$*, $m = 1, ..., M$, $k = 1, ..., K$.*
then $\sqrt{N}\big(\tilde{F}_{m_k;N}(B_k) - F_{m_k}(B_k)\big)_{k=1,...,K} \xrightarrow{d} \zeta \simeq \mathcal{N}(\mathbb{O}_K, \Sigma)$ as $N \to \infty$, where $\tilde{F}_{m_k;N}$ are the weighted empirical distributions defined in (14)*.*

Proof 1. Consistency of the estimators is obtained by lemma 1 from [12]. Theorem 2 from [3] states the asymptotic normality of the functional moments, and of the necessary estimators as the partial case.

2. Theorem 2.3.1 from [5] states that $\sup_{x\in\mathbb{R}}\sqrt{N}\big|\tilde{F}_{m_k}(x) - \hat{F}_{m_k}(x)\big| \xrightarrow{P} 0$ as $N \to \infty$, $k = 1, ..., K$. From here, and from point 1, obtain the necessary statement. □

We define the vectors of indicators

$$t_{j;N} := \big(\mathbb{I}_{\xi_{j;N}\in A_q}\big)_{q=1,...,Q} \in \mathbb{R}^Q. \tag{17}$$

Then the estimate of f^i takes form

$$\hat{f}_N^i := \frac{1}{N}\sum_{j=1}^{N} a_{j;N}^i t_{j;N},\ i = 1, 2. \tag{18}$$

To test the H_0 we will use the test statistic $\hat{T}_N$. One should reject the null hypothesis if $\hat{T}_N$

differs from zero significantly, and accept otherwise. Statistic $\hat{T}_N$ for hypothesis (4) takes form

$$\begin{aligned}\hat{T}_N &:= \hat{f}^1_N - f^0 \\ &= \frac{1}{N}\sum_{j=1}^{N} a^1_{j;N} t_{j;N} - f^0.\end{aligned} \tag{19}$$

And for hypothesis (5) $\hat{T}_N$ becomes

$$\begin{aligned}\hat{T}_N &:= \hat{f}^2_N - \hat{f}^1_N \\ &= \frac{1}{N}\sum_{j=1}^{N} (a^2_{j;N} - a^1_{j;N}) t_{j;N}.\end{aligned} \tag{20}$$

3 Asymptotics of the test statistics

Assume that statistic $\hat{T}_N$ takes form

$$\hat{T}_N = \frac{1}{N}\sum_{j=1}^{N} b_{j;N} t_{j;N} - t_0, \tag{21}$$

where $b_{j;N} \in \mathbb{R}$ is some set of weight coefficients, $t_0 \in \mathbb{R}^Q$ is some nonrandom vector. Both statistics (19) and (20) take this form.

Tests for checking the hypotheses (4) and (5) by statistic (21) are modifications of more general scheme for the tests introduced in [3].

Let us introduce the formal random values

$$\zeta_m := \left(\mathbb{I}_{\eta_m \in A_q}\right)_{q=1,\dots,Q}, \; m = 1, \dots, M. \tag{22}$$

Denote their covariance matrix as

$$\begin{aligned}\Phi_m &:= \operatorname{Var} \zeta_m \\ &= \left(f^m_{r,s} - f^m_r f^m_s\right)_{r,s=1,\dots,Q}.\end{aligned} \tag{23}$$

If the sets A_q, $q = 1, \dots, Q$ are disjoint, Φ_m takes form

$$\Phi_m = \operatorname{diag}[f^m] - f^m (f^m)^T, \tag{24}$$

where $f^m := (f^m_q)_{q=1,\dots,Q} \in \mathbb{R}^Q$ is the vector from f^m_q. In this case

$$\det[\Phi_m] = \Big(\prod_{q=1}^{Q} f^m_q\Big)\Big(1 - \sum_{q=1}^{Q} f^m_q\Big). \tag{25}$$

Denote the covariance matrix of $\hat{T}_N$ multiplied by the number of observations as

$$D_N := N \cdot \operatorname{Var}[\hat{T}_N] \; \in \mathbb{R}^{Q\times Q}. \tag{26}$$

The limit matrix for D_N we denote as

$$D := \lim_{N\to\infty} D_N \; \in \mathbb{R}^{Q\times Q}. \tag{27}$$

Lemma 1. *Let $A_1, \dots, A_Q$ be some measurable sets from $\mathfrak{F}$, and test statistic $\hat{T}_N$ takes form* (21)*. Then*
(i) Elements of D_N are expressed as

$$\begin{aligned}(D_N)_{k,l} &= \sum_{m=1}^{M} f^m_{k,l} \langle (b_{;N})^2 p^m_{;N} \rangle_N \\ &- \sum_{r,s=1}^{M} f^s_k f^s_l \langle (b_{;N})^2 p^r_{;N} p^s_{;N} \rangle_N.\end{aligned} \tag{28}$$

(ii) If the limit values $\langle (b)^2 p^m \rangle$ and $\langle (b)^2 p^r p^s \rangle$ exist, then exists the limit matrix D, and it's elements are expressed as

$$(D)_{k,l} = \sum_{m=1}^{M} f^m_{k,l} \langle (b)^2 p^m \rangle - \sum_{r,s=1}^{M} f^s_k f^s_l \langle (b)^2 p^r p^s \rangle. \tag{29}$$

Proof (i) Since $t_{j;N}$ are independent, $\operatorname{Var}[\hat{T}_N] = \frac{1}{N^2}\sum_{j=1}^{N} (b_{j;N})^2 \operatorname{Var}[t_{j;N}]$. At the same time, $\operatorname{Var}[t_{j;N}]_{k,l} = \sum_{m=1}^{M} p^m_{j;N} f^m_{k,l} - \sum_{r,s=1}^{M} p^r_{j;N} p^s_{j;N} f^r_k f^s_l$, $k, l = 1, \dots, Q$. Note that the last expression always exist.
(ii) Follows from (i). □

To formulate the following statements we need to find the conditions under which the matrices (26) and (27) are non-singular. They are positively definite since it is some covariance matrices. Thus, to matrices (26) and (27) be non-singular, it is sufficient to obtain some their lower estimates. Lower estimate is meant in Loewner sense. I.e. for matrices A and B we write $A \geq B$ if the matrix $(A - B)$ is non-negatively definite. Lemma 2 gives us the needed estimate.

Lemma 2. *Let $A_1, \dots, A_Q$ be some measurable sets from $\mathfrak{F}$, and test statistic $\hat{T}_N$ takes form* (21)*. Then the following fulfills.*
(i) $D_N \geq Z_N$, where

$$Z_N := \sum_{m=1}^{M} \langle (b_{;N})^2 p^m_{;N} \rangle_N \Phi_m \; \in \mathbb{R}^{Q\times Q}. \tag{30}$$

(ii) If the limits $\langle (b)^2 p^m \rangle$ and $\langle (b)^2 p^r p^s \rangle$ exist, then $D \geq Z$, where

$$Z := \sum_{m=1}^{M} \langle (b)^2 p^m \rangle \Phi_m \; \in \mathbb{R}^{Q\times Q}. \tag{31}$$

Proof (i) Let ρ_1, ρ_2 be some square-integrable random values. By the total variance law $\mathrm{Var}[\rho_1] = \mathbf{E}[\mathrm{Var}[\rho_1|\rho_2]] + \mathrm{Var}[\mathbf{E}[\rho_1|\rho_2]]$. I.e. $\mathrm{Var}[\rho_1] \geq \mathbf{E}[\mathrm{Var}[\rho_1|\rho_2]]$.
$\delta^m_{j;N} := \mathbb{I}_{O_{j;N} \in \mathcal{P}_m}$. Denote $\delta^m_{j;N} := \mathbb{I}_{\{O_{j;N} \in \mathcal{P}_m\}}$. Let $c \in \mathbb{R}^Q$ be any non-random vector. Then
$\mathrm{Var}[c^T \hat{T}_N] \geq \mathbf{E}[\mathrm{Var}\left[c^T \hat{T}_N \middle| \{\delta^m_{j;N}\}\right]]$
$= \mathbf{E}[\mathrm{Var}\left[\sum_{m=1}^M c^T \zeta_m \frac{1}{N} \sum_{j=1}^N b_{j;N} \delta^m_{j;N} \middle| \{\delta^m_{j;N}\}\right]]$
$= \mathbf{E}\left[\sum_{m=1}^M (\frac{1}{N} \sum_{j=1}^N b_{j;N} \delta^m_{j;N})^2 \, \mathrm{Var}[c^T \zeta_m]\right]$
$= \frac{1}{N} \sum_{m=1}^M \langle (b_{;N})^2 p^m_{;N} \rangle_N c^T \Phi_m c = \frac{1}{N} c^T Z_N c$.
Thus, $D_N \geq Z_N$.
(ii) Follows from (i) and lemma 1. □

The next theorem shows the asymptotic normality of $\hat{T}_N$.

Theorem 2. *Let $A_1, ..., A_Q$ be some measurable sets from $\mathfrak{F}$, and test statistic $\hat{T}_N$ takes form* (21). *Assume that*
(i) Weight coefficients $b_{j;N}$ take form $b_{j;N} = \sum_{i=1}^d h_i a^{m_i}_{j;N}$, and $t_0 = \sum_{i=1}^d h_i f^{m_i}$, where $h_1, ..., h_d$ is some set of real numbers.
(ii) Matrix Γ exists, is finite, and $\det \Gamma \neq 0$.
(iii) Values $\langle (b)^2 p^m \rangle$, $\langle (b)^2 p^r p^s \rangle$ exist, and are finite, $m, r, s = 1, ..., M$.
(iv) Exists $m_0 \in \{1, ..., M\}$ such that $\langle (b)^2 p^{m_0} \rangle > 0$.
(v) Matrix Φ_{m_0} is positively defined.
Then, under H_0, $\sqrt{N}(\hat{T}_N - t_0) \xrightarrow{d} \zeta \simeq \mathcal{N}(\mathbb{O}_Q, D)$. Moreover, matrix D is positively defined.
If conditions (iii) and (iv) of theorem 1 are fulfilled (for $K = d$), then the statement remains true for $b_{j;N} = \sum_{i=1}^d h_i \tilde{a}^{m_i}_{j;N}$.

Proof From conditions (iii), (iv), (v) and lemma 2 it follows that D exists and is positively definite. From conditions (i), (ii) and theorem 1 obtain consistency and asymptotic normality for $\hat{T}_N$. □

Remark 1. Weight coefficients $b_{j;N}$ for statistics (19) and (20) are the partial case of coefficients $b_{j;N} = \sum_{i=1}^d h_i a^{m_i}_{j;N}$. Indeed, for (19) we need to set $d = 1$, $h_1 = 1$, $m_1 = 1$, and for (20): $d = 2$, $h_1 = -1$, $h_2 = 1$, $m_1 = 1$, $m_2 = 2$.

Remark 2. If the sets $A_1, ..., A_Q$ are disjoint, then condition (v) of theorem 2 can be replaced by condition
(v') $f^{m_0}_q > 0$, $q = 1, ..., Q$, $\sum_{q=1}^Q f^{m_0}_q < 1$.
Indeed, from condition (v') and equality (25) obtain $\det \Phi_{m_0} > 0$. Since Φ_{m_0} is the covariance matrix, it is positively definite.

4 Test construction

Theorem 2 shows that the statistic $\hat{T}_N$ of the form (21) has the mean value t_0, and is asymptotically normal with the dispersion matrix D. Note that the matrix D is non-singular. It's elements can be estimated as

$$(\hat{D}_N)_{k,l} := \sum_{m=1}^M \hat{f}^m_{k,l;N} \langle (b_{;N})^2 p^m_{;N} \rangle_N - \sum_{r,s=1}^M \hat{f}^r_{k;N} \hat{f}^s_{l;N} \langle (b_{;N})^2 p^r_{;N} p^s_{;N} \rangle_N. \qquad (32)$$

Thus, we can expect that under H_0 statistic

$$\hat{s}_N := N(\hat{T}_N - t_0)^T \hat{D}_N^{-1} (\hat{T}_N - t_0) \qquad (33)$$

converges by distribution to χ^2 distribution with Q degrees of freedom (next we will write χ^2_Q). Since matrix D is non-singular, and its estimate $\hat{D}_N$ is consistent (under conditions from theorem 1), then the matrix $\hat{D}_N$ is non-singular for large enough N.

Remark 3. To estimate D in hypothesis (4), we can put more precise estimate (28) instead of (32). Indeed, if (4) fulfills, then we already know the values f^m_k and $f^m_{k,l}$.

Assume that $\hat{s}_N$ converges by distribution to χ^2_Q. For a given significance level α, we can test the hypothesis by comparing $\hat{s}_N$ with the threshold level:

$$\pi_{\alpha;N} : \begin{cases} \text{accept } H_0 & \text{if } \hat{s}_N \leq Q^{\chi^2_Q}(1-\alpha), \\ \text{reject } H_0 & \text{otherwise,} \end{cases} \qquad (34)$$

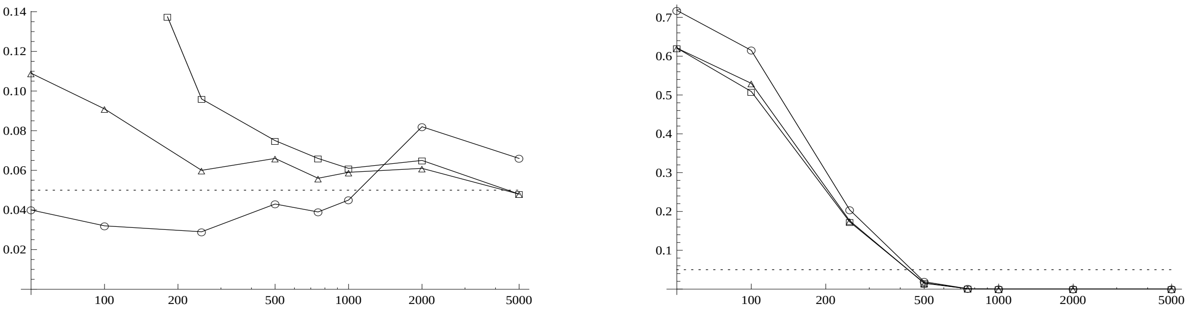
where $Q^{\chi^2_Q}(1-\alpha)$ is the quantile of level $1-\alpha$ for the distribution χ^2_Q. The reached significance level (p-level) can be calculated as

$$p := 1 - F_{\chi^2_Q}(\hat{s}_N), \qquad (35)$$

where $F_{\chi^2_Q}$ is the distribution function of χ^2_Q.

Theorem 3 proves that the test (34) is defined correctly. For weight coefficients $b_{j;N}$ defined through $a^m_{j;N}$, this theorem is the direct consequence of theorem 5 from [3]. If we use improved coefficients $\tilde{a}^m_{j;N}$, then it's proof is the same as the proof of the mentioned theorem.

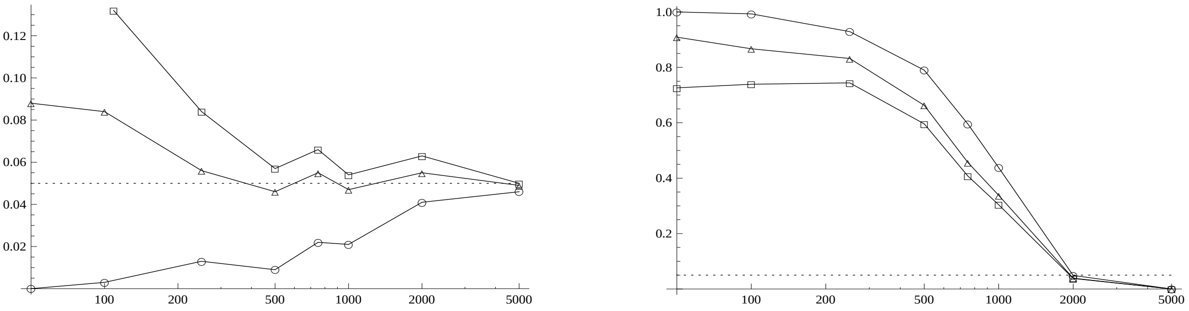
Theorem 3. *Assume the conditions of theorem 2 are fulfilled. Then, under H_0, $\lim_{N\to\infty} \mathbf{P}[\pi_{\alpha;N}$ rejects $H_0] = \alpha$.*



A1. First-type errors (H_0 is true)

A2. Second-type errors (H_0 is false)

Рис. 1: Frequences of errors. Test checks if the distribution is equal to some known in advance.



B1. First-type errors (H_0 is true)

B2. Second-type errors (H_0 is false)

Рис. 2: Frequences of errors. Test checks if the distribution of first two components are equal.

5 Numerical results

Introduced approach was tested by the simulation study. We took three-component mixture. Each component has the Gaussian distribution with some mean μ_m and variance σ^2_m, $m = 1, 2, 3$. The concentrations were generated with the uniform distribution: $p_{j;N} = (u_1, u_2, u_3)^T/s$, where $u_i \simeq \mathrm{Unif}(0, 1)$, $i = 1, 2, 3$, $s := u_1+u_2+u_3$.

In each experiment it was generated 1000 samples with 50, 100, 250, 500, 750, 1000, 2000 and 5000 observations. As the result, we measured the frequences of the first- and second-type errors for significance level $\alpha = 0.05$.

It was implemented three modifications of $\pi_{\alpha;N}$. In the first modification (next (ss)) both estimates $\hat{T}_N$ and $\hat{D}_N$ were constructed via minimax weight coefficients $a^m_{j;N}$ from (11). In the second modification (next (si)) we built $\hat{T}_N$ via coefficients $a^m_{j;N}$, but we used improved coefficients $\tilde{a}^m_{j;N}$ from (14) to estimate $\hat{D}_N$. The third modification (next (ii)) represents $\hat{T}_N$ and $\hat{D}_N$ both constructed via improved weights.

The results of all three realizations are shown on figures 1 and 2. Line marked by □ represents the realizations (ss), marked by △ – (si), and by ○ – (ii). The dashed line is for the significance level $\alpha = 0.05$.

Experiments A. For this pair of experiments we took the Gaussian mixture, which components have zero values $\mu_1 = 0$, $\mu_2 = 1$, $\mu_3 = 0$ and variances $\sigma^2_1 = 1$, $\sigma^2_2 = 4$, $\sigma^2_3 = 9$. In experiment A1 the distribution of the first component was compared to the Gaussian distribution with parameters $\mu_0 = 0$ and $\sigma^2_0 = 1$, i.e. when H_0 is true. We defined the sets A_q as the intervals $\big(\mu_0 + (i - \frac{1}{2})\sigma_0, \mu_0 + (i + \frac{1}{2})\sigma_0\big]$, $i = -2, ..., 2$. Frequencies of the first-type errors are shown on the left part of figure 1. Analogously, in experiment A2 we took the Gaussian distribution with paramenters $\mu_0 = 1$ and $\sigma_0 = 1$, i.e. H_0 is false. Frequencies of the second-type errors are shown on the right part of figure 1.

Experiments B. For this pair of experiments, we checked the hypothesis about equality of the first two components' distributions. We took the sets A_q analogously to experiment A, except that μ_0 and σ_0 were estimated as $\tilde{\mu}_{0;N} :=$

$\frac{1}{2}(\tilde{\mu}_{1;N} + \tilde{\mu}_{2;N})$ and $\tilde{\sigma}^2_{0;N} := \frac{1}{2}(\tilde{\sigma}^2_{1;N} + \tilde{\sigma}^2_{2;N})$ respectfully, where $\tilde{\mu}_{m;N} := \frac{1}{N}\sum_{j=1}^{N} \tilde{a}^m_{j;N}\xi_{j;N}$, $\tilde{\sigma}^2_{m;N} := \frac{1}{N}\sum_{j=1}^{N} \tilde{a}^m_{j;N}(\xi_{j;N} - \tilde{\mu}_{m;N})^2$, $m = 1, 2$. In experiment B1 we took mean values as $\mu_1 = \mu_2 = 0$, $\mu_3 = 3$, and variances as $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_3^2 = 9$. I.e. H_0 is true. Frequences of the first-type errors are shown on figure 2. In experiment B2 the mean values were taken as $\mu_1 = \mu_2 = \mu_3 = 0$, and the variances as $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\sigma_3^2 = 9$. I.e. H_0 is false. Frequences of the second-type errors are shown on figure 2.

Список використаних джерел

[1] Доронін, О. В. Нижня межа матриці розсіяння для семіпараметричного оцінювання у моделі суміші /О. В. Доронін // Теорія ймовірностей та математична статистика. — 2014.— Вип. 90. — С. 64–76.

[2] Доронін, О. В. Адаптивне оцінювання у семіпараметричній моделі суміші, /О. В. Доронін // Теорія ймовірностей та математична статистика. — 2014.— Вип. 91. — С. 27–38.

[3] Doronin, A., Maiboroda, R. Testing hypotheses on moments by observations from a mixture with varying concentrations, /A. Doronin, R. Maiboroda // Modern Stochastics: Theory and Applications. — 2014.— Вип. 1, №2. — С. 61–70.

[4] Лодатко, А., Майборода, Р. Адаптивна моментна оцінка параметру розподілу по спостереженнях з домішкою, /А. Лодатко, Р. Майборода // Теорія ймовірностей та математична статистика. — 2006.— Вип. 75. — С. 61–70.

[5] Майборода, Р. Є., Сугакова, О. В. Оцінювання та класифікація за спостереженнями із суміші, /Р. Є. Майборода, О. В. Сугакова // Київ: Київський університет. — 2008.

[6] Autin, F., Pouet, Ch. Test on the components of mixture densities, /F. Autin, Ch. Pouet // Statistics & Risk Modelling. — 2011.— Вип. 28, №4. — С. 389–410.

6 Conclusion

The model of mixture with varying concentrations was considered. We developed the technique of test construction to check the hypotheses about equality of components' distributions. This technique can easily be extended to the case of more than two mixture's components. Quality of the tests was checked by the simulation study. Developed tests can be applied to the analysis of statistical data from medical, biological, sociological, political, economical etc. areas.

References

1. DORONIN, O.V. (2014) Lower bound of dispersion matrix for semiparametric estimation in mixture model. *Theory of Probability and Mathematical Statistics.* 90, p. 64–76.

2. DORONIN, O.V. (2014) Adaptive estimation in semiparametric model of mixture with varying concentrations. *Theory of Probability and Mathematical Statistics.* 91, p. 27–38.

3. DORONIN, O. and MAIBORODA, R. (2014) Testing hypotheses on moments by observations from a mixture with varying concentrations. *Modern Stochastics: Theory and Applications.* 1, 2, p. 195–209.

4. LODATKO, N. and MAIBORODA, R. (2007) An adaptive moment estimator of a parameter of a distribution constructed from observations with admixture. *Theory of Probability and Mathematical Statistics.* 75, p. 71–82.

5. MAIBORODA, R. and SUGAKOVA, O. (2008) *Estimation and Classification of Observations from Mixtures.* Kyiv: Kyiv University Publishers.

6. AUTIN, F. and POUET, Ch. (2011) Test on the components of mixture densities. *Statistics & Risk Modelling.* 28, 4, p. 389–410.

[7] Bordes, L., Delmas, C., Vandekerkhove, P. Semiparametric Estimation of a two-component Mixture model where one component is known, /L. Bordes, C. Delmas, P. Vandekerkhove // Scandinavian Journal of Statistics. — 2006.— Вип. 33. — С. 733–752.

[8] Hall, P., Zhou, X.-H. Nonparametric estimation of component distributions in a multivariable mixture, /P. Hall, X.-H. Zhou // Annals of Statistics. — 2003.— Вип. 31, №1. — С. 201–224.

[9] Maiboroda, R. E., Kubaichuk, O. O. Improved estimators for moments constructed from observations of a mixture, /R. E. Maiboroda, O. O. Kubaichuk // Theory of Probability and Mathematical Statistics. — 2005.— Вип. 70. — С. 83–92.

[10] Maiboroda, R., Sugakova, R. Nonparametric density estimation for symmetric distributions by contaminated data, /R. Maiboroda, O. Sugakova // Metrica. — 2012.— Вип. 75, №1. — С. 109–126.

[11] Maiboroda, R., Sugakova, O. Statistics of mixtures with varying concentrations with application to DNA microarray data analysis, /R. Maiboroda, O. Sugakova // Journal of nonparametric statistics. — 2012.— Вип. 24, №1. — С. 201–205.

[12] Maiboroda, R. E., Sugakova, O. V., Doronin, A. V. Generalized estimating equations for mixtures with varying concentrations, /R. E. Maiboroda, O. V. Sugakova, A. V. Doronin // The Canadian Journal of Statistics. — 2013.— Вип. 41, №2. — С. 217–236.

[13] McLachlan, G. J., Peel, D. Finite Mixture Models, /G. J. McLachlan, D. Peel // New York: Wiley.— 2000.

[14] Sugakova, O. Adaptive estimates for the parameter of a mixture of two symmetric distributions, /O. Sugakova // Theory of Probability and Mathematical Statistics. — 2011.— Вип. 82. — С. 149–159.

[15] Titterington, D. M., Smith, A. F., Makov, U. E. Analysis of Finite Mixture Distributions, /D. M. Titterington, A. F. Smith, U. E. Makov // New York: Wiley.— 1985.

7. BORDES, L., DELMAS, C. and VANDEKERKHOVE, P. (2006) Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics.* 33, p. 733–752.

8. HALL, P. and ZHOU, X.-H. (2003) Nonparametric estimation of component distributions in a multivariable mixture. *Annals of Statistics.* 31, 1, p. 201–224.

9. MAIBORODA, R.E. and KUBAICHUK, O.O. (2005) Improved estimators for moments constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics.* 70, p. 83–92.

10. MAIBORODA, R. and SUGAKOVA, O. (2012) Nonparametric density estimation for symmetric distributions by contaminated data. *Metrica.* 75, 1, p. 109–126.

11. MAIBORODA, R. and SUGAKOVA, O. (2012) Statistics of mixtures with varying concentrations with application to DNA microarray data analysis. *Journal of nonparametric statistics.* 24, 1, p. 201–205.

12. MAIBORODA, R.E., SUGAKOVA, O.V. and DORONIN A.V. (2013) Generalized estimating equations for mixtures with varying concentrations *The Canadian Journal of Statistics.* 41, 2, p. 217–236.

13. MCLACHLAN, G.J. and PEEL, D. (2000) *Finite Mixture Models.* New York: Wiley.

14. SUGAKOVA, O. (2011) Adaptive estimates for the parameter of a mixture of two symmetric distributions. *Theory of Probability and Mathematical Statistics.* 82, p. 149–159.

15. TITTERINGTON, D.M., SMITH, A.F. and MAKOV, U.E. (1985) *Analysis of Finite Mixture Distributions.* New York: Wiley.

Received: 06.06.2014