MSC 90C35, 90C27 DOI: 10.14529/mmp180202

# MODELLING THE FLOW OF CHARACTER RECOGNITION RESULTS IN VIDEO STREAM

*V.V. Arlazarov*[1,2], *O.A. Slavin*[1,2], *A.V. Uskov*[1], *I.M. Janiszewski*[1]
[1]Federal Research Center "Informatics and Management" of RAS, Moscow, Russian Federation
[2]JSC "Smart Engines Service", Moscow, Russian Federation
E-mail: bvva777@gmail.com, oslavin@isa.ru, uskov@isa.ru, yanishevsky@isa.ru

The paper considers problems of developing stochastic models consistent with results of character image recognition in video stream. A set of assumptions that define the models structure and properties is stated. A class of distributions, namely the Dirichlet distribution and its generalizations, that set a description of the model components is pointed out; and methods for statistical estimation of the distribution parameters are given. To rank the models, the Akaike information criterion is used. The proposed theoretical distributions are verified vs sample data.

*Keywords: stochastic model; video stream; the character recognition; Dirichlet distribution; Akaike criterion; goodness-of-fit Anderson-Darling tests.*

## Introduction

Recently there has been a steadily growing interest in document management systems built on mobile platforms. An integral part of this kind of systems is automatic document input subsystems in which digital cameras mobile devices and web cameras act as a "scanning" device. Obvious problems [1–4], arising in the process of shooting a document made by a camera-mobile device, and subsequently at the stages of image processing, including the process of its recognition. Currently, these document images are of a lower quality than those obtained on the scanning device. Therefore, to obtain accurate and reliable recognition results, along with the use of traditional methods of document image recognition, it is necessary to develop new methods based on the processing of a single video stream as a digital image of the document. With this approach, there are several new problems, among which the following can be noted. The first is the problem of assessing the necessary volume of observations to make a decision on reliable recognition of a single symbol or field of the document. The second is the determination of the final estimates of the recognition results based on the integration of partial resolution of the document recognition and its fields on each frame of the video stream. To solve the integration problem, there are quite numerous nonparametric methods [1, 2, 5]. The most of them are based on the use of elementary statistics.

In this paper we consider the problem of construction and investigation of stochastic models describing the results of document recognition. The use of the proposed models, in our opinion, will allow us to solve the above problems productively.

We will consider the recognition of structured documents consisting of a set of text fields with pre-known properties [1, 3, 6]. For such fields not only tracking of sequence of recognition results of fields, but analysis of a sequence of recognition results of one symbol is possible [7].

## 1. Problem Statement

Let there be a sequence of frames $\{I^k\}_{k=1}^K$ for some document. For each $I^k$ there is a field $F(I^k)$, without loss of generality consisting of character space $A^k$. We assume that character space $A^k$ is a set of $n$ alternatives, namely $\{\langle s_i, X_i^k\rangle\}_{i=1}^n$, where $s_i$ is the character code of the Cyrillic alphabet $Z$ ($s_1 =$'А', $s_2 =$'Б', etc.), $X_i^k$ is the probability alternative (classifier estimate) obtained at the detection of frame $I^k$. Introduce a notation for vectors of estimates $\mathbf{X}^k = (X_1^k, \ldots, X_n^k)^T$. Let $\mathbf{X}^k \in \mathbb{T}^n$, where $\mathbb{T}^n$ is a simplex

$$\mathbb{T}^n = \{(X_1, \ldots, X_n)^T : \; X_i \geq 0, \, i = 1, \ldots, n; \, \sum_{i=1}^n X_i = 1\}. \tag{1}$$

In addition, we assume that the recognition results $\mathbf{X}^k$, $k = 1, \ldots, K$, is a sample of independent equally distributed values. Let's call the sequence $\{\mathbf{X}^k\}_{k=1}^K$ the recognition results flow.

The paper deals with the problems of modelling (approximation) of the empirical distribution of the recognition results flow $\{\mathbf{X}^k\}_{k=1}^K$. It is possible to distinguish four stages of solving the problem [8]:

1) model choice, i.e. hypothesize affiliation families of distributions;

2) estimate parameters;

3) evaluate quality of fit;

4) estimate goodness of fit statistical tests.

In the classical formulation of the modelling problem based on the existing sample of independent random variables with unknown distribution density belonging to a certain family of parametric distributions, it is required to construct estimates of unknown parameters using the maximum likelihood principle. This problem may not have a solution if the dimension of the parameter vector is large and far exceeds the sample volume. Next, a concept is proposed that allows viewing a parametric family as a combination of distributions with vectors of smaller dimension parameters. This approach allows obtaining parameter estimates for small sample volumes of recognition results.

## 2. Preliminary Observations and Properties

Here are the main symbols, as well as some definitions and results from [9]. Consider two positive random vectors $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ associated with $X_i = \frac{Y_i}{Y^+}$, where $Y^+ = \sum_{i=1}^n Y_i$. In literature the vector $\mathbf{X}$ is called compositional data, and the vector $\mathbf{Y}$ is called basis. Note that the estimates $\mathbf{X}^i$ can be considered as an example of compositional data.

Quite often the components can be grouped according to some homogeneity criterion. In such cases, it is of interest to study the totals and relative values within each group. In order to formalize this approach, it is accepted to use amalgamation and subcomposition, which can be explained as follows. Let $a_0 = 0 < a_1 < \ldots < a_{C-1} < a_C = n$ be the set of indexes and

$$X_1, \ldots, X_{a_1} | X_{a_1+1}, \ldots, X_{a_2} | \ldots | X_{a_{c-1}+1}, \ldots, X_{a_c} \tag{2}$$

be a complete partition (of order $c-1$) of subsets of the vector $\mathbf{X}$. Based on partition (2), we define the subcomposition with index $i$:

$$\mathbf{S}_i = (X_{a_{i-1}+1}, \ldots, X_{a_i}) / X_i^+, \tag{3}$$

where $X_i^+ = X_{a_{i-1}+1} + \ldots + X_{a_i}$, $i = 1, \ldots, c$ . The amalgamation is the vector of the totals of the $c$ subsets $X^+ = (X_1^+, \ldots, X_c^+)$.

It is known [10] that for modelling of composite data the presence of the composite invariance property is essential. The basis of $Y$ is compositionally invariant if the corresponding composition $X = C(Y)$ is independent of $Y^+$. In fact, all versions of the notions of independence presented in the literature can be expressed in terms of subcompositions $\mathbf{S}_i$, $i = 1, \ldots, c$, and amalgamation $X^+$. For example, consider the most popular partition case of order 1 ($c = 2$). We denote independence by $\perp$ and a set of independent random variables by $A \perp B \perp C$. Let $a_1 = m$. Then, partition independence means that $\mathbf{S}_1 \perp \mathbf{S}_2 \perp X^+$; subcompositional invariance means that $(\mathbf{S}_1, \mathbf{S}_2) \perp X^+$; neutrality on the left means that $\mathbf{S}_1 \perp (\mathbf{S}_2, X^+)$; neutrality on the right means that $\mathbf{S}_2 \perp (\mathbf{S}_1, X^+)$; subcompositional independence means that $\mathbf{S}_1 \perp \mathbf{S}_2$.

## 3. Dirichlet Distribution and Its Generalizations

Dirichlet distribution is one of the key multidimensional distributions for composite data modelling. It plays an important role for the representation of proportions. This distribution has a simple form and has many convenient mathematical properties [11]. However, the Dirichlet distribution is considered to be insufficiently flexible. Therefore, generalizations of Dirichlet distribution were proposed by different authors [9, 10].

A random vector $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{T}^n$ has a Dirichlet distribution if the distribution density is as follows

$$f_D(\mathbf{x}_n; \alpha) = \frac{\Gamma(\alpha_+)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^n x_i^{\alpha_i - 1}, \tag{4}$$

where $\alpha$ is the vector of positive parameters, $\alpha_+ = \sum_{i=1}^n \alpha_i$.

There is a simple relation [11] between the parameters of the joint density and the marginal densities of each component $X_i \sim Beta(\alpha_i, \alpha_+ - \alpha_i)$.

Let's make the following transformation

$$X_1^* = X_1,\ X_i^* = \frac{X_i}{(1 - \sum\limits_{j=1}^{i-1} X_j)},\ i = 2, \ldots, n-1. \tag{5}$$

Then for these random variables is fair

$$X_1^* \sim Beta(\alpha_1, \sum_{j=2}^n \alpha_j) \text{ and } X_i^* | \, X_1, \ldots, X_{i-1} \sim Beta(\alpha_i, \sum_{j=i+1}^n \alpha_j),\ i = 2, \ldots, n-1. \tag{6}$$

The flexible Dirichlet distribution $FD^n(\alpha, \mathbf{p}, \tau)$ was first proposed in [9]. Let $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{T}^n$. The distribution function of the vector $\mathbf{X} \sim FD^n(\alpha, \mathbf{p}, \tau)$ is a finite mixture of Dirichlet distributions

$$FD^n(\mathbf{x}, \alpha, \mathbf{p}, \tau) = \sum_{i=1}^n p_i D^n(\mathbf{x}; \alpha + \tau \mathbf{e}_i), \tag{7}$$

where $\mathbf{e}_i$ is a vector whose elements are all equal to zero except for the $i$-th element which is equal to one, and the density of the distribution is as follows

$$f_{FD}(\mathbf{x};\alpha,\mathbf{p},\tau)=\frac{\Gamma(\alpha_+ +\tau)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\left(\prod_{i=1}^{n}x_i^{\alpha_i-1}\right)\left(\sum_{i=1}^{n}p_i\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i+\tau)}x_i^{\tau}\right), \tag{8}$$

where $\mathbf{x}\in\mathbb{T}^n$; $i=1,\ldots,n$; $\alpha_i>0$, $\alpha_+=\sum_{i=1}^{n}\alpha_i$; $0\le p_i<1$, $\sum_{i=1}^{n}p_i=1$; $\tau>0$.

The marginal distributions of vector $\mathbf{X}$ components can be represented as follows

$$X_i\sim p_iBeta(\alpha_i+\tau,\alpha_+-\alpha_i)+(1-p_i)Beta(\alpha_i,\alpha_+-\alpha_i+\tau),\ i=1,\ldots,n. \tag{9}$$

A vector $\mathbf{X}=(X_1,\ldots,X_n)^T\in\mathbb{T}^n$ is said to follow a Connor – Mosimann distribution $CM(\alpha,\beta)$ if the density can be represented as follows [12]:

$$f_{CM}(\mathbf{x};\alpha,\beta)=[\prod_{i=1}^{n-1}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)}x_i^{\alpha_i-1}(\sum_{j=i}^{n}x_j)^{\beta_{i-1}-(\alpha_i+\beta_i)}]x_n^{\beta_{n-1}-1}, \tag{10}$$

where $\mathbf{x}\in\mathbb{T}^n$; $\alpha_i>0$, $i=1,\ldots,n$, $\beta_j>0$, $j=1,\ldots,n-1$, $\beta_0=0$. Apply a transformation similar to (5)

$$X_1^*=X_1,\ X_i^*=\frac{X_i}{(1-\sum\limits_{j=1}^{i-1}X_j)},\ i=2,\ldots,n-1. \tag{11}$$

Accordingly, conditional distributions are defined as follows

$$X_1^*\sim Beta(\alpha_1,\beta_1)\,\text{and}\,X_i^*|\,X_1,\ldots,X_{i-1}\sim Beta(\alpha_i,\beta_i),\ i=2,\ldots,n-1. \tag{12}$$

The Dirichlet distribution is closely related to the parametric family of multivariate Liouville distributions. For consideration take from this family the beta-Liouville distribution. Let the vector $\mathbf{X}\in(0,1)^n$ have a stochastic representation $\mathbf{X}=R\mathbf{Y}$, where $R\perp\mathbf{Y}$, $R=\sum_{i=1}^{n}X_i$, $R\sim Beta(a,b)$, $\mathbf{Y}=(Y_1,\ldots,Y_n)^T\in\mathbb{T}^n$, $\mathbf{Y}\sim Dir(\alpha)$. A random vector $\mathbf{X}$ is said to follow a beta-Liouville distribution. The density of beta-Liouville distribution has the form [13]:

$$f_{BL}(x_1,\ldots,x_n;a,b,\alpha_1,\ldots,\alpha_n)=$$
$$=\frac{\Gamma(a+b)\Gamma\left(\sum\limits_{i=1}^{n}\alpha_i\right)}{\Gamma(a)\Gamma(b)\prod\limits_{i=1}^{n}\Gamma(\alpha_i)}\times\left(\sum_{i=1}^{n}x_i\right)^{a-\sum\limits_{i=1}^{n}\alpha_i-1}\left(1-\sum_{i=1}^{n}x_i\right)^{b-1}\prod_{i=1}^{n}x_i^{\alpha_i-1}. \tag{13}$$

## 4. Mathematical Model

Consider the vector $\mathbf{X}\in\mathbb{T}^n$. Without loss of generality, assume that $X_1\ge\ldots\ge X_n$. Let's set some criteria, using which we can divide the composition $\mathbf{X}$ into two subcompositions $\mathbf{X}^{(1)}=(X_1,\ldots,X_m)^T$ and $\mathbf{X}^{(2)}=(X_{m+1},\ldots,X_n)^T$. For example, the first subcomposition includes are all elements whose values exceed a certain level $\mathcal{L}$, and all the rest elements are in the second. Both subcompositions may be represented as follows

$$\mathbf{X}^{(1)} = X_{+}^{(1)} \frac{\mathbf{X}^{(1)}}{X_{+}^{(1)}}, \quad \mathbf{X}^{(2)} = X_{+}^{(2)} \frac{\mathbf{X}^{(2)}}{X_{+}^{(2)}}, \tag{14}$$

where $X_{+}^{(1)} = \sum_{i=1}^{m} X_i$, $X_{+}^{(2)} = \sum_{i=m+1}^{n} X_i$, $X_{+}^{(2)} = 1 - X_{+}^{(1)}$. Next, introduce new variables

$$\mathbf{Z}^{(1)} = \frac{\mathbf{X}^{(1)}}{X_{+}^{(1)}}, \quad \mathbf{Z}^{(2)} = \frac{\mathbf{X}^{(2)}}{1 - X_{+}^{(1)}}, \quad R = X_{+}^{(1)}. \tag{15}$$

It is possible to write an expression

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{pmatrix} = \begin{pmatrix} R \cdot \mathbf{Z}^{(1)} \\ (1 - R) \cdot \mathbf{Z}^{(2)} \end{pmatrix}, \tag{16}$$

where $\mathbf{X}^{(1)}$ is an $m$-dimensional vector; $\mathbf{X}^{(2)}$ is an $(n-m)$-dimensional vector.

Formulate assumptions for the variables included in the right part of (16).

*Assumption 1.* Let $\mathbf{Z}^{(1)} \perp \mathbf{Z}^{(2)} \perp R$.

*Assumption 2.* Let $R \sim Beta(a, b)$.

*Assumption 3.* Let $\mathbf{Z}^{(2)} \sim Dir(\alpha^{(2)})$.

*Assumption 4.* Let $\mathbf{Z}^{(1)} \sim Dir(\alpha^{(1)})$.

*Assumption 5.* Let $\mathbf{Z}^{(1)} \sim FD(\alpha^{(1)}, \mathbf{p}, \tau)$.

*Assumption 6.* Let $\mathbf{Z}^{(1)} \sim CM(\alpha^{(1)}, \beta^{(1)})$.

For simplification of designations consider

$$\alpha = \begin{pmatrix} \alpha^{(1)} \\ \alpha^{(2)} \end{pmatrix} = (\alpha_1, \ldots, \alpha_n). \tag{17}$$

where $\alpha^{(1)}$ is an $m$-dimensional vector of parameters; $\alpha^{(2)}$ is an $(n-m)$-dimensional vector of parameters.

Using the assumptions, we construct three stochastic models for the distribution of composition $\mathbf{X}$.

*Model 1.* If assumptions 1 – 4 are satisfied, the density of the composition $\mathbf{X}$ has the form

$$\begin{gathered} f_1(x_1, \ldots, x_n; a, b, \alpha_1, \ldots, \alpha_n) = \\ = \frac{\Gamma(a+b)\Gamma\left(\sum\limits_{i=1}^{m} \alpha_i\right)\Gamma\left(\sum\limits_{i=m+1}^{n} \alpha_i\right)}{\Gamma(a)\Gamma(b)\prod\limits_{i=1}^{n}\Gamma(\alpha_i)} \times \\ \times \left(\sum_{i=1}^{m} x_i\right)^{a - \sum\limits_{i=1}^{m} \alpha_i - 1} \left(\sum_{i=m+1}^{n} x_i\right)^{b - \sum\limits_{i=m+1}^{n} \alpha_i - 1} \prod_{i=1}^{n} x_i^{\alpha_i - 1}. \end{gathered} \tag{18}$$

Indeed, let us write density of $(\mathbf{Z}^{(1)}, R, \mathbf{Z}^{(2)})$

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} r^{a-1}(1-r)^{b-1} \times \frac{\Gamma(\sum\limits_{i=1}^{m} \alpha_i)}{\prod\limits_{i=1}^{m}\Gamma(\alpha_i)} \prod_{i=1}^{m} [z_i^{(1)}]^{\alpha_i - 1} \times \frac{\Gamma(\sum\limits_{i=m+1}^{n} \alpha_i)}{\prod\limits_{i=m+1}^{n}\Gamma(\alpha_i)} \prod_{i=m+1}^{n} [z_i^{(2)}]^{\alpha_i - 1}. \tag{19}$$

Now consider $x_1, \ldots, x_n, x_+^{(1)}$. The corresponding Jacobian is given by

$$\mathbb{J}(z_1^{(1)}, \ldots, z_m^{(1)}, r, z_{m+1}^{(2)}, \ldots, z_n^{(2)} \to x_1, \ldots, x_n, x_+^{(1)}) = (x_+^{(1)})^{-m}(1 - x_+^{(1)})^{-(n-m)}. \tag{20}$$

Thus, from (15), (19) and (20) we have (18).

*Model 2.* If assumptions 1 – 3, 5 are satisfied, the density of the composition **X** has the form

$$\begin{gathered} f_2(x_1, \ldots, x_n; a, b, \alpha_1, \ldots, \alpha_n, p_1, \ldots, p_m, \tau) = \\ = \frac{\Gamma(a+b)\Gamma\left(\sum\limits_{i=1}^{m}\alpha_i + \tau\right)\Gamma\left(\sum\limits_{i=m+1}^{n}\alpha_i\right)}{\Gamma(a)\Gamma(b)\prod\limits_{i=1}^{n}\Gamma(\alpha_i)} \left(\sum_{i=1}^{m} x_i\right)^{a - \sum\limits_{i=1}^{m}\alpha_i - \tau - 1} \times \\ \times \left(\sum_{i=m+1}^{n} x_i\right)^{b - \sum\limits_{i=m+1}^{n}\alpha_i - 1} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \left(\sum_{i=1}^{m} p_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \tau)} x_i^{\tau}\right). \end{gathered} \tag{21}$$

*Model 3.* If assumptions 1 – 3, 6 are satisfied, the density of the composition **X** has the form

$$\begin{gathered} f_3(x_1, \ldots, x_n; a, b, \alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_{m-1}) = \\ = \frac{\Gamma(a+b)\Gamma\left(\sum\limits_{i=m+1}^{n}\alpha_i\right)}{\Gamma(a)\Gamma(b)\prod\limits_{i=m+1}^{n}\Gamma(\alpha_i)} \left(\sum_{i=1}^{m} x_i\right)^{a-1} \left(\sum_{i=m+1}^{n} x_i\right)^{b - \sum\limits_{i=m+1}^{n}\alpha_i - 1} \times \\ \times \prod_{i=1}^{m-1}\left[\frac{\Gamma(\alpha_i + \beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x_i^{\alpha_i - 1}\left(\sum_{j=i}^{m} x_j\right)^{\beta_{i-1} - (\alpha_i + \beta_i)}\right] x_m^{\beta_{m-1} - 1} \prod_{i=m+1}^{n} x_i^{\alpha_i - 1}. \end{gathered} \tag{22}$$

## 5. Estimation of Model Parameters

For estimating the Dirichlet parameter vector, the principle of maximum likelihood is usually used. We assume that the parameter values that provide the maximum of the log-likelihood function are taken as estimates:

$$\begin{gathered} L(\mathbf{x}_n^{(j)}; \alpha) = \sum_{j=1}^{k} \log f_D(\mathbf{x}_n^{(j)}; \alpha) = \\ = k\left\{\log\Gamma(\sum_{i=1}^{n}\alpha_i) - \sum_{i=1}^{n}\log\Gamma(\alpha_i) + \sum_{i=1}^{n}(\alpha_i - 1)\log G_i\right\}, \end{gathered} \tag{23}$$

where $G_i = (\prod\limits_{j=1}^{n} x_{ji})^{\frac{1}{k}},\ i = 1, \ldots, n.$

It is known [14] that the function $L$ is globally concave, since the Dirichlet distribution belongs to the exponential family, and the Newton – Raphson algorithm converges to the global optimum [11, 15].

Estimation of the parameters of a flexible Dirichlet distribution is considered as the problem of separating a finite mixture of Dirichlet distributions, for the solution of which the EM algorithm [16, 17] can be suitably adapted. Suppose we have $k$ independent observations $\mathbf{x}_j$, $j = 1, \ldots, k$, each having a distribution (8). Further complete vector data $\mathbf{x}_c$ is given by:

$$\mathbf{x}_c = (\mathbf{x}, \mathbf{v}) = (\mathbf{x}_1, \mathbf{v}_1, \ldots, \mathbf{x}_k, \mathbf{v}_k), \tag{24}$$

where vector $\mathbf{v}_j = (v_{j1}, \ldots, v_{jn})$ represents the missing data with $v_{ji}$ being equal to 1 if the $j$-th observation has arisen from the $i$-th component of the mixture model and 0 otherwise.

The log-likelihood function with respect to (7) and (24) has the form

$$\log L_c(\theta) = \sum_{j=1}^{k} \sum_{i=1}^{n} v_{ji} [\log p_i + \log f_D(\mathbf{x}_j; \alpha + \tau \mathbf{e}_i)], \tag{25}$$

where $\theta = (\alpha, \mathbf{p}, \tau)$; $f_D(\mathbf{x}_j; \alpha + \tau \mathbf{e}_i)$ is the Dirichlet density.

The $s + 1$ step of the EM algorithm can be described as follows.

**E-step**: given the current parameter estimates $\theta^{(s)} = (\alpha^{(s)}, \mathbf{p}^{(s)}, \tau^{(s)})$ and $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$, calculate the conditional expectation of the complete-data log-likelihood

$$Q(\theta; \theta^{(s)}) = \sum_{j=1}^{k} \sum_{i=1}^{n} p_i(\mathbf{x}_j; \theta^{(s)}) [\log p_i + \log f_D(\mathbf{x}_j; \alpha + \tau \mathbf{e}_i)], \tag{26}$$

where $p_i(\mathbf{x}_j; \theta^{(s)})$ represents "posterior" probability that $\mathbf{x}_j$ belongs to the $i$-th component of the mixture given $\theta^{(s)}$ which is defined as follows

$$p_i(\mathbf{x}_j; \theta) = \frac{p_i f_D(\mathbf{x}_j; \alpha + \tau \mathbf{e}_i)}{\sum_{r=1}^{n} p_r f_D(\mathbf{x}_j; \alpha + \tau \mathbf{e}_r)}, \; i = 1, \ldots, n. \tag{27}$$

**M-step**: maximize (26) to obtain the maximum likelihood estimates of $\theta^{(s+1)}$

$$\theta^{(s+1)} = \arg\max_{\theta} Q(\theta; \theta^{(s)}). \tag{28}$$

In particular, we have $p_i^{(s+1)} = \frac{1}{k} \sum_{j=1}^{k} p_i(\mathbf{x}_j; \theta^{(s)})$, $i = 1, \ldots, n-1$, whereas $\alpha^{(s+1)}, \tau^{(s+1)}$ can be computed by implementing a Newton – Raphson method. Iterating occurs until the "sufficiently small" change of the observed log-likelihood (or the parameter estimates) is reached.

Estimates of the parameters of the Connor – Mosimann distribution can be obtained using properties (12) [12]. Perform the estimation of the parameters of the beta distribution, $\alpha_r$, $\beta_r$ $(r = 1, \ldots, n-1)$, and take the resulting estimates for estimates of the prior distribution.

## 6. Applying Criteria to Rank a Model

In this section, we consider the problem of selection a model from a set of competing models, which gives the best in the sense of not approaching the characteristic of the studied flow of recognition results. In modern statistics analysis for the purposes of ranking models uses a simple and effective tool – the Akaike information criterion (AIC), which can be represented as follows

$$Q_{AIC}(F^{(k)}, \hat{\theta}^{(k)}) = -2 \sum_{j=1}^{n} \ln f^{(k)}(\mathbf{y}_j; \hat{\theta}^{(k)}) + 2q^{(k)}, \tag{29}$$

where $F^{(k)}$ is the $k$-th model; $f^{(k)}(\cdot)$ is the distribution density for $F^{(k)}$; $\mathbf{y}_j$ is the observation, $j = 1, \ldots, n$; $\hat{\theta}^{(k)}$ is the vector-parameter for $F^{(k)}$; $q^{(k)}$ is the number of parameters on which $F^{(k)}$ depends.

The choice of the model consists in ranking the models in accordance with the values of the $Q_{AIC}$ and the preference of the model with its lowest value.

**Remark 1.** Based on the AIC, we can construct a partition rule for the composition $\mathbf{X}$. Suppose that for some set of dimensions of subcompositions $\mathcal{M} = \{m_{\min}, \ldots, m_{\max}\}$, assumptions 1 – 3 hold. Then the required dimension for, for example, model 1 can be defined as the solution of the problem

$$m^* = \arg\min_{m \in \mathcal{M}} (-2 \sum_{k=1}^{K} \ln f_1(X_1^k, \ldots, X_n^k; \hat{a}, \hat{b}, \hat{\alpha}_1, \ldots, \hat{\alpha}_m, 1, \ldots, 1) + 2m). \tag{30}$$

It is easy to see that the obtained results (18), (21), and (22) differ in the type of distribution that describes the value of $\frac{\mathbf{X}^{(1)}}{X_+^{(1)}}$. Therefore, we confine ourselves to calculating the values of the partial criterion $Q_{AIC}$ (Table 1).

**Table 1**

| Models | $Q_{AIC}$ |
|---|---|
| Model 1 (Dirichlet distribution) | −221,73 |
| Model 2 ( flexible Dirichlet distribution) | −228,50 |
| Model 3 (Connor – Mosimann distribution) | −234,39 |

## 7. Test of Models Fit

In this section, we consider the question of how well the proposed models are consistent with the observations. Before proceeding to the construction of objective quantitative estimates, it is useful to subject the obtained models to the procedure of informal graphic diagnostics, that is, to compare the sample data with the parametric model by graphic methods. To visualize the quality-of-fit of models, we use the so-called "Q-Q (quantile-quantile) plot", which is a method for comparing the empirical and the theoretical distributions by plotting their quantiles against each other. If the theoretical distribution is well matched, the points on the plot are located along a straight line. Figs. 1 to 4 show the graphs for the proposed models.

Before proceeding to the formulation and testing of the hypotheses of interest, we give the formula of "standardization" $X \sim Beta(a, b)$

$$X^* = I(X; a, b) = \frac{B(X; a, b)}{B(a, b)} = \frac{1}{B(a, b)} \int_0^X t^{a-1}(1-t)^{b-1} dt, \tag{31}$$

where $I(z, a, b)$ is the regularized incomplete beta function; $B(z, a, b)$ is the incomplete beta function; $B(a, b)$ is the beta function. It is known [18] that the transformed random variable has the following property

$$X^* \sim Beta(1, 1). \tag{32}$$

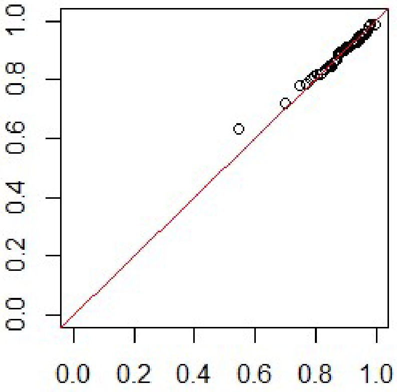

**Fig. 1**. The Q-Q plot for the variable $X_{+}^{(1)}$ in the verification of assumption 2

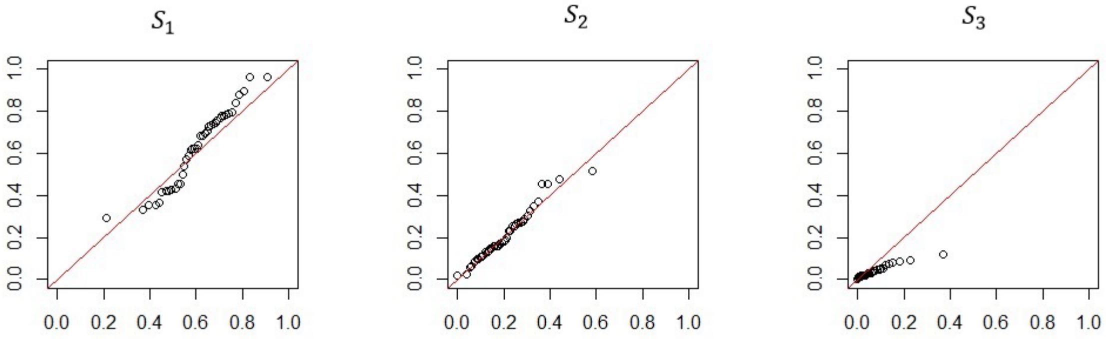

**Fig. 2**. The Q-Q plot for the variable $\mathbf{X}^{(1)}$ in model 1

We need a similar to (31) transformation formula for $Y$ corresponding to a mixture of beta distributions

$$Y \sim p \cdot Beta(a_1, a_2) + (1 - p) \cdot Beta(b_1, b_2). \tag{33}$$

Introduce a transformation

$$Y^* = p \cdot I(Y; a_1, a_2) + (1 - p) \cdot I(Y; b_1, b_2), \tag{34}$$

then $Y^* \sim Beta(1, 1)$.

Formally, all hypotheses of goodness-of-fit necessary for our purposes have a general form. Let $X_1, \ldots, X_n$ be a sequence of independent identically distributed random variables with distribution $F$. Then the null hypothesis is

$$H_0 : \ \ F(x) = I(x; 1, 1). \tag{35}$$

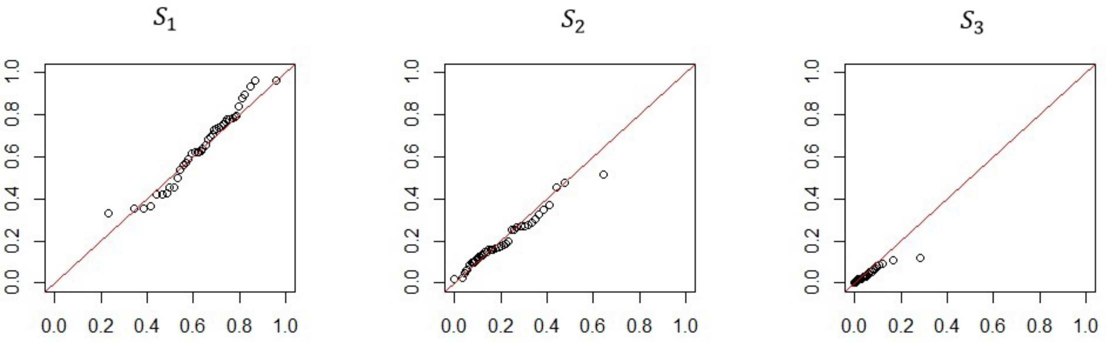

**Fig. 3**. The Q-Q plot for the variable $\mathbf{X}^{(1)}$ in model 2

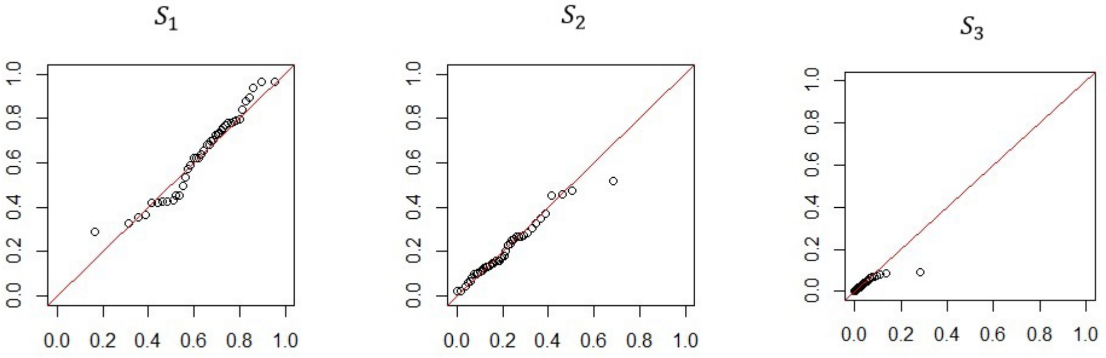

**Fig. 4**. The Q-Q plot for the variable $\mathbf{X}^{(1)}$ in model 3

To verify the correspondence between the sample distribution and the theoretical law, we use the Anderson–Darling tests of goodness-of-fit [19], based on statistic [20]

$$S_\Omega = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i,\theta) + \left(1-\frac{2i-1}{2n}\right)\ln(1-F(x_i,\theta))\right\}. \tag{36}$$

We point out that large values of $S_\Omega$ statistics indicate poor compliance. The distribution of the statistics $S_\Omega$ rapidly approaches the asymptotic distribution, which has the form [21]

$$\begin{aligned} a2(S) = \frac{\sqrt{2\pi}}{S}\sum_{j=0}^{\infty}(-1)^j\frac{\Gamma(j+\frac{1}{2})(4j+1)}{\Gamma(\frac{1}{2})\Gamma(j+1)}\exp\left\{-\frac{(4j+1)^2\pi^2}{8S}\right\}\times \\ \times\int_0^\infty \exp\left\{\frac{S}{8(y^2+1)} - \frac{(4j+1)^2\pi^2y^2}{8S}\right\}dy. \end{aligned} \tag{37}$$

For practical purposes, this distribution may be used provided that the sample size is greater than 5 [19].

In conclusion, consider the hypothesis testing scheme. We take assumption 4. Since $\mathbf{Z}^{(1)} \sim Dir(\alpha^{(1)})$, using properties (5) and (6) of the Dirichlet distribution, one can obtain

$$\begin{gathered} Z_1^* = Z_1^{(1)}, \quad Z_i^* = \frac{Z_i^{(1)}}{(1-\sum\limits_{j=1}^{i-1} Z_j^{(1)})} \\ Z_1^* \sim Beta(\alpha_1^{(1)}, \sum\limits_{j=2}^{m} \alpha_j^{(1)}), \\ Z_i^* |\, Z_1^{(1)}, \ldots, Z_{i-1}^{(1)} \sim Beta(\alpha_i^{(1)}, \sum\limits_{j=i+1}^{m} \alpha_j^{(1)}),\ i = 2, \ldots, m-1. \end{gathered} \tag{38}$$

Apply the transformation (31) to$Z_i^*$ and obtain

$$\begin{gathered} Y_1 = I(Z_1^*; \alpha_1^{(1)}, \sum\limits_{j=2}^{m} \alpha_j^{(1)}), \\ Y_i = I(Z_i^*; \alpha_i^{(1)}, \sum\limits_{j=i+1}^{m} \alpha_j^{(1)}),\ i = 1, \ldots, m-1. \end{gathered} \tag{39}$$

Then, given (32), we can formulate the hypothesis of goodness-of-fit in the following form

$$H_0:\ F_i(y) = I(y; 1, 1), \tag{40}$$

where $F_i(y)$ is the distribution function of $Y_i$, $i = 1, \ldots, m-1$; $I(y, 1, 1)$ is the regularized incomplete beta function with parameters (1,1).

**Table 2**

The results of the goodness-of-fit test

| Alternative | Model | $S_\Omega$ | $p$-value |
|---|---|---|---|
| <K> | 1 | 1,4679 | 0,1844 |
| | 2 | 0,67315 | 0,5807 |
| | 3 | 0,51872 | 0,7273 |
| <E> | 1 | 1,0981 | 0,3095 |
| | 2 | 0,51691 | 0,7286 |
| | 3 | 0,13964 | 0,9992 |

As it can be seen, when setting the significance level $\alpha < 0,18$, there is no cause for rejecting the hypotheses tested by the goodness-of-fit test for all models.

## Conclusion

In this paper we propose new probabilistic models describing the results of document recognition in video stream. The concept of the flow of recognition results is introduced. The considered models suggest that the result of sign recognition in the field of the document can be represented as a combination of random variables and random vectors. For various assumptions, expressions for the density of the results of sign recognition are obtained. Methods for parameters estimation are presented. The Akaike information criterion is used for ranking models. Tests that confirmed the adequacy of the stochastic models were carried out. In conclusion, note that the solution of the problem of integration the parameters obtained by simulation of the flow of the recognition results can be used for [7].

**Acknowledgement.** *This work is supported by Russian Foundation for Basic Research (projects No. 16-07-01051 and 17-29-03297).*

## References

1. Hartl A., Arth C., Schmalstieg D. Real-time Detection and Recognition of Machine-Readable Zones with Mobile Devices. *Proceedings 10th International Conference on Computer Vision Theory and Applications (VISAPP 2015)*, 2015, pp. 79–87. DOI: 10.5220/0005294700790087
2. Tian S., Yin X.C., Su Y., Hao H.W. Unified Framework for Tracking Based Text Detection and Recognition from Web Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, vol. 40, no. 3, pp. 542–554. DOI: 10.1109/TPAMI.2017.2692763
3. Arlazarov V.V., Zhukovsky A.E., Krivtsov V.E., Nikolaev D.P., Polevoy D.V. [Analysis of the Features of Using Stationary and Mobile Small-Sized Digital Video Cameras for Document Recognition]. *Information Technology and Computer Systems*, 2014, no. 3, pp. 71–81. (in Russian)
4. Bulatov K., Arlazarov V., Chernova T.V., Slavin A., Nikolaev D. Smart IDReader: Document Recognition in a Video Stream. *The 14th IAPR International conference on document Analysis and Recognition (ICDAR 2017), master classes and lessons: November 9-12, Kyoto, Japan*, 2017, pp. 39–44.
5. Bulatov K.B., Kirsanov V., Arlazarov V. et al. Methods for Integrating the Results of Recognition of Document Text Fields in the Video Stream of a Mobile Device. *RFBR Journal*, 2016, no. 4, pp. 109–115. (in Russian) DOI: 10.22204/2410-4639-2016-092-04-109-115
6. Arlazarov V.L., Marchenko A.E., Sholomov D.L. Cumulative Contexts in the Recognition Problem. *Proceedings of the Institute of Systems Analysis, Russian Academy of Sciences (ISA RAS)*, 2014, vol. 64, no. 4, pp. 64–72. (in Russian)
7. Bulatov K.B. Choosing the Optimal Strategy for Combining Frame-by-Frame Character Recognition Results in a Video Stream. *Information Technology and Computer Systems*, 2017, no. 3, pp. 45–55. (in Russian)
8. Ricci V. *Fitting Distributions with R.* 2005, 24 p. Available at: https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf
9. Ongaro A., Migliorati S. A Generalization of the Dirichlet Distribution. *Journal of Multivariate Analysis*, 2013, vol. 114, pp. 412–426. DOI: 10.1016/j.jmva.2012.07.007
10. Connor R, Mosimann J. Concepts of Independence for Proportions with a Generalisation of the Dirichlet Distribution. *Journal of the American Statistical Association*, 1969, vol. 64, no. 325, pp. 194–206. DOI: 10.1080/01621459.1969.10500963
11. Ng K.W., Tian G.-L., Tang M.-L. *Dirichlet and Related Distributions: Theory, Methods and Applications.* Chichester, Wiley, 2011. DOI: 10.1002/9781119995784
12. Elfadaly F, Garthwaite P. Obtaining Preliminary Dirichlet and Connor – Mosimann Distributions for Polynomial Models. *Test*, 2013, vol. 22, no. 4, pp. 628–646. DOI: 10.1007/s11749-013-0336-4

13. Fang K., Kotz S., Ng K.W. *Symmetric Multivariate and Related distribution.* New York, Chapman and Hall, 1989.
14. Ronning G. Maximum Likelihood Estimation of Dirichlet Distributions. *Journal of Statistical Computation and Simulation*, 1989, vol. 32, no. 3, pp. 215–221. DOI: 10.1080/00949658908811178
15. Robitzsch A. *Sirt: Supplementary Item Response Theory Models. R Package Version 2.6-9.* Avialable at: https://cran.r-project.org/web/packages/sirt/index.html
16. Migliorati S., Ongaro A., Monti G.S. A Structured Dirichlet Mixture Model for Compositional Data: Inferential and Applicative Issues. *Statistics and Computing*, 2016, vol. 27, no. 4, pp. 963–983. DOI: 10.1007/s11222-016-9665-y
17. Migliorati C., A. Di Brisco M., Vestrucci M. *FlexDir: Tools to Work with the Flexible Dirichlet Distribution. Package R version 1.0.* Avialable at: https://cran.r-project.org/web/packages/FlexDir/index.html
18. Li Y. *Goodness-of-Fit Tests for Dirichlet Distributions with Applications: PhD Thesis.* Bowling Green State University, 2015.
19. Stephens M.A. Goodness of Fit, Anderson–Darling Test. *Encyclopedia of Statistical Sciences*, 2006. 4 p. DOI: 10.1002/0471667196.ess0041.pub2
20. Lemeshko B.Yu., Lemeshko S.B., Postavalov S.N., Chimitova E.V. *Statisticheskiy analiz dannykh, modelirovanie i issledovanie veroyatnostnykh zakonomernostey. Kompyuternyy podkhod* [Statistics Data Analysis, Simulation and Study of Probabilistic Regularities]. Novosibirsk, Infra-M, 2011. (in Russian)
21. Bolshev L.N., Smirnov N.V. *Tablicy matematicheskoy statistiki* [Tables of Mathematical Statistics]. Moscow, Nauka, 1983. (in Russian)

*Received April 20, 2018*

**УДК 004.021** **DOI: 10.14529/mmp180202**

# МОДЕЛИРОВАНИЕ ПОТОКА РЕЗУЛЬТАТОВ РАСПОЗНАВАНИЯ СИМВОЛОВ В ВИДЕОПОСЛЕДОВАТЕЛЬНОСТЯХ

***В.В. Арлазаров***[1,2]***, О.А. Славин***[1,2]***, А.В. Усков***[1]***, И.М. Янишевский***[1]

[1]Федеральный исследовательский центр «Информатика и управление» РАН,
г. Москва, Российская Федерация
[2]ООО «Смарт Энджинс Сервис», г. Москва, Российская Федерация

В данной работе рассматриваются проблемы построения вероятностных моделей, согласованных с результатами распознавания образов символов в видеопоследовательностях. Сформулирована совокупность предположений, определяющих структуру и свойства построенных моделей. Выделен класс распределений, а именно распределение Дирихле и его обобщения, задающих описание компонентов моделей, и приведены методы статистического оценивания параметров указанных распределений. Для ранжирования моделей используется информационный критерий Акаике. Проведена проверка согласия предложенных теоретических распределений выборочным данным.

*Ключевые слова: вероятностная модель; видеопоследовательность; распознавание символов; распределение Дирихле; критерий Акаике; критерий согласия Андерсона – Дарлинга.*

## Литература

1. Hartl, A. Real-Time Detection and Recognition of Machine-Readable Zones with Mobile Devices / A. Hartl, C. Arth, D. Schmalstieg // Proceedings 10th International Conference on Computer Vision Theory and Applications (VISAPP 2015). – 2015. – P. 79–87.
2. Tian, S. Unified Framework for Tracking Based Text Detection and Recognition from Web Videos / S. Tian , X.C. Yin, Y. Su, H.W. Hao // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2018. – V. 40, № 3. – P. 542–554.
3. Арлазаров, В.В. Анализ особенностей использования стационарных и мобильных малоразмерных цифровых видеокамер для распознавания документов / В.В. Арлазаров, А.Е. Жуковский, В.Е. Кривцов, Д.П. Николаев, Д.В. Полевой // Информационные технологии и вычислительные системы. – 2014. – № 3. – С. 71–81.
4. Bulatov, K. Smart IDReader: Document Recognition in Video Stream / K. Bulatov, V. Arlazarov, T. Chernov, O. Slavin, D. Nikolaev // The 14th IAPR International Conference on Document Analysis and Recognition (ICDAR 2017). – 2017. – P. 39–44.
5. Булатов, К. Методы интеграции результатов распознавания текстовых полей документов в видеопотоке мобильного устройства / К. Булатов, В. Кирсанов, В.В. Арлазаров и др. // Вестник РФФИ. – 2016. – № 4. – С. 109–115.
6. Арлазаров, В.Л. Накопительные контексты в задаче распознавания / В.Л. Арлазаров, А.Е. Марченко, Д.Л. Шоломов // Труды ИСА РАН. – 2014. – Т. 64, № 4. – С. 64–72.
7. Булатов, К.Б. Выбор оптимальной стратегии комбинирования покадровых результатов распознавания символа в видеопотоке / К.Б. Булатов// Информационные технологии и вычислительные системы. – 2017. – № 3. – С. 45–55.
8. Ricci, V. Fitting Distributions with R / V. Ricci. – 2005. – 24 p. – URL: https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf
9. Ongaro, A. Generalization of the Dirichlet Distribution / A. Ongaro, S.A. Migliorati // Journal of Multivariate Analysis. – 2013. – V. 114. – P. 412–426.
10. Connor, R. Concepts of Independence for Proportions with a Generalisation of the Dirichlet Distribution / R. Connor, J.J. Mosimann // Journal of the American Statistical Association. – 1969. – V. 64, № 325. – P. 194–206.
11. Ng, K.W. Dirichlet and Related Distributions: Theory, Methods and Applications /K.W. Ng, G.-L. Tian, M.-L. Tang. – Chichester: Wiley, 2011.
12. Elfadaly, F. Eliciting Dirichlet and Connor – Mosimann Prior Distributions for Multinomial Models / F. Elfadaly, P. Garthwaite // Test. – 2013. – V. 22, № 4. – P. 628–646.
13. Fang, K. Symmetric Multivariate and Related Distributions / K. Fang, S. Kotz, K.W. Ng. – N.Y.: Chapman and Hall, 1990.
14. Ronning, G. Maximum Likelihood Estimation of Dirichlet Distributions / G. Ronning // Journal of Statistical Computation and Simulation. – 1989. – V. 32, № 3. – P. 215–221.
15. Robitzsch, A. Sirt: Supplementary Item Response Theory Models. R Package Version 2.6-9 / A. Robitzsch. – URL: https://cran.r-project.org/web/packages/sirt/index.html
16. Migliorati, S. A Structured Dirichlet Mixture Model for Compositional Data: Inferential And Applicative Issue / S. Migliorati, A. Ongaro, G.S. Monti // Statistics and Computing. – 2017. – V. 27, № 4. – P. 963–983.
17. Migliorati, S. FlexDir: Tools to Work with the Flexible Dirichlet Distribution. R Package Version 1.0 / S. Migliorati, A.M. Di Brisco, M. Vestrucci. – URL: https://cran.r-project.org/web/packages/FlexDir/index.html

18. Li, Y. Goodness-of-Fit Tests for Dirichlet Distributions with Applications: PhD Thesis / Y. Li. – Bowling Green State University, 2015.
19. Stephens, M.A. Goodness of Fit, Anderson-Darling Test of / M.A. Stephens// Encyclopedia of Statistical Sciences. – 2006. – 4 p.
20. Лемешко, Б.Ю. Статистический анализ данных, моделирование и исследование вероятностных закономерностей / Б.Ю. Лемешко, С.Б. Лемешко, С.Н. Постовалов, Е.В. Чимитова. – М.: НИЦ ИНФРА-М, 2015.
21. Большев, Л.Н. Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов. – М.: Наука, 1983.

Владимир Викторович Арлазаров, кандидат технических наук, заведующий отделом, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация); ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация), bvva777@gmail.com.

Олег Анатольевич Славин, доктор технических наук, главный научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация); ООО «Смарт Энджинс Сервис» (г. Москва, Российская Федерация), oslavin@isa.ru.

Анатолий Васильевич Усков, кандидат физико-математических наук, заведующий лабораторией, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация), uskov@isa.ru.

Игорь Михайлович Янишевский, кандидат физико-математических наук, старший научный сотрудник, Федеральный исследовательский центр «Информатика и управление» РАН (г. Москва, Российская Федерация), yanishevsky@isa.ru.

*Поступила в редакцию 20 апреля 2018 г.*