Stochastic Analysis and Applications, 24: 61–78, 2006
Copyright © Taylor & Francis, LLC
ISSN 0736-2994 print/1532-9356 online
DOI: 10.1080/07362990500397533

On Application of the Partition Distance Concept to a Comparative Analysis of Psychological or Sociological Tests

A. D'yachkov
Department of Probability Theory, Faculty of Mechanics and Mathematics, Moscow State University, Moscow, Russia

V. Rykov
Department of Mathematics, University of Nebraska at Omaha, Omaha, Nebraska, USA

D. Torney
Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

S. Yekhanin
Computer Science and Artificial Intelligence Laboratory, Cambridge, Massachusetts, USA

Abstract: We discuss two distance concepts between q-ary n-sequences, $2 \leq q < n$, called partition distances. This distances are metrics in the space of all partitions of a finite n-set. For the metrics, we study codes called q-partition codes and present a construction of these codes based on the first order Reed–Muller codes. A random coding bound is obtained. We also work out an application of q-partition codes to the statistical analysis of psychological or medical tests using questionnaires.

Keywords: Clusters analysis; Medical tests; Partition codes; Partitions distance; Partitions of sets; Psychological tests.

Received August 31, 2004; Accepted September 7, 2004
Address correspondence to V. Rykov, Department of Mathematics, University of Nebraska at Omaha, 6001 Dodge St., Omaha, Nebraska 68182-0243, USA; E-mail: vrykov@mail.unomaha.edu

Mathematics Subject Classification: Primary 62P15; Secondary 05A18, 94B60.

1. INTRODUCTION

The symbol $\triangleq$ denotes definitional equalities and the symbol $[n] \triangleq \{1, 2, \ldots, n\}$ denotes the set of integers from 1 to n.

Let $n > q \geq 2$ be fixed integers, $A_q \triangleq \{0, 1, \ldots, q-1\}$ be the standard q-ary alphabet, and $\mathscr{M}_q = \{\mu\}$, $\mu = \mu(x)$ be the set of all $q!$ one-to-one mappings (permutations) of A_q, i.e.,

$$y = \mu(x), \qquad x = \mu^{-1}(y), \quad x, y \in A_q = \{0, 1, \ldots, q-1\}, \ \ \mu, \mu^{-1} \in \mathscr{M}_q.$$

We will say that an arbitrary fixed q-ary n-sequence $\mathbf{x} \triangleq (x_1, x_2, \ldots, x_n) \in A_q^n$, identifies an unordered *q-partition* $\tilde{\mathbf{x}} \triangleq \{E_0(\mathbf{x}), E_1(\mathbf{x}), \ldots, E_{q-1}(\mathbf{x})\}$ of the set

$$[n] = E_0(\mathbf{x}) + E_1(\mathbf{x}) + \cdots + E_{q-1}(\mathbf{x}), \quad \text{where } E_x(\mathbf{x}) \triangleq \{i : x_i = x\}, \ \ x \in A_q.$$

Any $\tilde{\mathbf{x}}$ contains q', $1 \leq q' \leq q$, nonempty parts. $\mathbf{x}^\mu \triangleq (\mu(x_1), \mu(x_2), \ldots, \mu(x_n)) \in A_q^n$, is called *$\mu$-complement of* $\mathbf{x}$. Any $\mathbf{x}^\mu$, $\mu \in \mathscr{M}_q$, identifies the same q-partition $\tilde{\mathbf{x}}$.

Given $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in A_q^n$, $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in A_q^n$, and arbitrary fixed elements $x, y \in A_q$, we will denote by symbol $n(x, y \,|\, \mathbf{x}, \mathbf{y})$ the number of positions i, $i = 1, 2, \ldots, n$, in which $x_i = x$ and $y_i = y$. Therefore, for any $\mathbf{x}, \mathbf{y} \in A_q^n$ and $x, y \in A_q$, we have

$$0 \leq n(x, y \,|\, \mathbf{x}, \mathbf{y}) \leq n, \quad \sum_{x \in A_q} \sum_{y \in A_q} n(x, y \,|\, \mathbf{x}, \mathbf{y}) = n.$$

So that

$$S(\mathbf{x}, \mathbf{y}) \triangleq \sum_{x \in A_q} n(x, x \,|\, \mathbf{x}, \mathbf{y}) \quad \text{and} \quad H(\mathbf{x}, \mathbf{y}) \triangleq n - \sum_{x \in A_q} n(x, x \,|\, \mathbf{x}, \mathbf{y})$$

can be called, respectively, Hamming *similarity* and *distance* between $\mathbf{x}$ and $\mathbf{y}$.

1.1. Partition Distance $\mathscr{P}$

Obviously, $H(\mathbf{x}, \mathbf{y}^\mu) = n - S(\mathbf{x}, \mathbf{y}^\mu) = n - \sum_{x \in A_q} n(x, \mu^{-1}(x) \,|\, \mathbf{x}, \mathbf{y})$. Other properties of $H(\mathbf{x}, \mathbf{y})$ and $S(\mathbf{x}, \mathbf{y})$ connected with the complement operation are presented in Proposition 1.

Proposition 1. 1. *For any* $\alpha \in \mathscr{M}_q$ *and* $\beta \in \mathscr{M}_q$, *the minimum*

$$\min_{\mu \in \mathscr{M}_q} H(\mathbf{x}^\alpha, \mathbf{y}^\mu) = \min_{\mu \in \mathscr{M}_q} H(\mathbf{x}^\mu, \mathbf{y}^\beta) = \min_{\mu \in \mathscr{M}_q} H(\mathbf{x}, \mathbf{y}^\mu)$$

$$= n - \max_{\mu \in \mathscr{M}_q} S(\mathbf{x}, \mathbf{y}^\mu) = n - \max_{\mu \in \mathscr{M}_q} \left\{ \sum_{x \in A_q} n(x, \mu(x) \mid \mathbf{x}, \mathbf{y}) \right\}$$

$$\triangleq \mathscr{P}(\mathbf{x}, \mathbf{y}) = \mathscr{P}(\mathbf{y}, \mathbf{x})$$

and the number $\mathscr{P}(\mathbf{x}, \mathbf{y}) \geq 0$, *where the sign of equality is achieved if and only if* $\mathbf{x}$ *and* $\mathbf{y}$ *identify the same unordered q-partition of the set* $[n]$.

2. *For any* $\mathbf{x}, \mathbf{y} \in A_q^n$, *the number*

$$\mathscr{P}(\mathbf{x}, \mathbf{y}) \leq \frac{q-1}{q} \cdot n,$$

where the sign of equality is achieved, for example, if $n = qs \cdot q$, $s = q, 2q, \ldots$ *and*

$$\mathbf{x} = (0, 0, \ldots, 0, \ldots, q-1, \ldots, q-1),$$
$$\mathbf{y} = (0, 1 \ldots, q-1, \ldots, 0, \ldots, q-1).$$

Proof. 1) We use the evident identity $H(\mathbf{x}^\beta, \mathbf{y}^\alpha) = H(\mathbf{x}, \mathbf{y}^{\beta^{-1}\alpha})$.

2) Consider q mappings $\mu_k = \mu_k(x)$, $\mu_k \in \mathscr{M}_q$, $k = 0, 1, 2, \ldots, q-1$, having the form:

$$\mu_0(x) \triangleq x, \qquad \mu_k(x) \triangleq \mu_0(x + k(\mathrm{mod}\, q)), \quad k = 1, 2, \ldots, q-1, \quad x \in A_q.$$

It is easy to understand that for any pair of q-ary n-sequences $(\mathbf{x}, \mathbf{y})$, the sum

$$\sum_{k=0}^{q-1} \sum_{x \in A_q} n(x, \mu_k(x) \mid \mathbf{x}, \mathbf{y}) = \sum_{x \in A_q} \sum_{y \in A_q} n(x, y \mid \mathbf{x}, \mathbf{y}) = n.$$

Hence, there exists an integer k such that the internal sum in the left-hand side is at least n/q. From definition of $\mathscr{P}(\mathbf{x}, \mathbf{y})$ it follows the second statement. Proposition 1 is proved.

Proposition 1 motivates Definition 1.

Definition 1.

$$\mathscr{P}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \mathscr{P}(\mathbf{x}, \mathbf{y}) \triangleq \min_{\mu \in \mathscr{M}_q} H(\mathbf{x}, \mathbf{y}^\mu) = \min_{\mu \in \mathscr{M}_q} H(\mathbf{x}^\mu, \mathbf{y}) = \mathscr{P}(\mathbf{y}, \mathbf{x}) = \mathscr{P}(\tilde{\mathbf{y}}, \tilde{\mathbf{x}})$$

is called a *partition distance* (or $\mathscr{P}$*-distance*) between q-partitions $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ of the set $[n]$.

Remark 1. The distance $\mathscr{P} = \mathscr{P}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ between q-partitions $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ of the set $[n]$ is equal to the minimum number of elements that must be deleted from $[n]$, so that two induced q-partitions ($\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ restricted to the remaining elements) are identical. The partition distance $\mathscr{P} = \mathscr{P}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is also equal to the minimum number of elements of $[n]$ that must be moved between subsets $E_x(\mathbf{x})$, $x \in A_q$, of q-partition $\tilde{\mathbf{x}} = \{E_0(\mathbf{x}), E_1(\mathbf{x}), \ldots, E_{q-1}(\mathbf{x})\}$, so that the resulting q-partition equals $\tilde{\mathbf{y}}$. Such form of the partition distance definition was suggested in [1].

Proposition 2. $\mathscr{P}(\mathbf{x}, \mathbf{y})$ *satisfies the triangle inequality*

$$\mathscr{P}(\mathbf{x}, \mathbf{y}) \le \mathscr{P}(\mathbf{x}, \mathbf{z}) + \mathscr{P}(\mathbf{z}, \mathbf{y}) \quad \textit{or} \quad \mathscr{P}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \le \mathscr{P}(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}) + \mathscr{P}(\tilde{\mathbf{z}}, \tilde{\mathbf{y}}).$$

Hence, $\mathscr{P}$-distance is a metric in the space of unordered q-partitions.

Proof. By Definition 1, we have $\mathscr{P}(\mathbf{x}, \mathbf{z}) + \mathscr{P}(\mathbf{z}, \mathbf{y}) = \min_{\mu \in \mathscr{M}_q} H(\mathbf{x}, \mathbf{z}^\mu) + \min_{\mu \in \mathscr{M}_q} H(\mathbf{z}, \mathbf{y}^\mu)$. Consider a mapping $\alpha \in \mathscr{M}_q$, for which $\min_{\mu \in \mathscr{M}_q} \times H(\mathbf{x}, \mathbf{z}^\mu) \triangleq H(\mathbf{x}, \mathbf{z}^\alpha)$. In virtue of Proposition 1, $\min_{\mu \in \mathscr{M}_q} H(\mathbf{z}, \mathbf{y}^\mu) = \min_{\mu \in \mathscr{M}_q} H(\mathbf{z}^\alpha, \mathbf{y}^\mu)$. Let for the given α, $\min_{\mu \in \mathscr{M}_q} H(\mathbf{z}^\alpha, \mathbf{y}^\mu) \triangleq H(\mathbf{z}^\alpha, \mathbf{y}^\tau)$. One can write $\mathscr{P}(\mathbf{x}, \mathbf{z}) + \mathscr{P}(\mathbf{z}, \mathbf{y}) = H(\mathbf{x}, \mathbf{z}^\alpha) + H(\mathbf{z}^\alpha, \mathbf{y}^\tau) \ge H(\mathbf{x}, \mathbf{y}^\tau) \ge \mathscr{P}(\mathbf{x}, \mathbf{y})$, where we apply the triangle inequality for Hamming distance and Definition 1. Proposition 2 is proved.

1.2. Partition Distance $\widetilde{\mathscr{P}}$

Another definition of partition distance was suggested in [2].

Definition 2 [2]**.** Partition distance $\widetilde{\mathscr{P}} = \widetilde{\mathscr{P}}(\mathbf{x}, \mathbf{y}) = \widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ (or $\widetilde{\mathscr{P}}$-distance) between q-partitions

$$\tilde{\mathbf{x}} = \{E_0(\mathbf{x}), E_1(\mathbf{x}), \ldots, E_{q-1}(\mathbf{x})\}, \quad E_x(\mathbf{x}) = \{i : x_i = x\}, \ \ x \in A_q,$$

and

$$\tilde{\mathbf{y}} = \{E_0(\mathbf{y}), E_1(\mathbf{y}), \ldots, E_{q-1}(\mathbf{y})\}, \quad E_y(\mathbf{y}) = \{i : y_i = y\}, \ \ y \in A_q,$$

is equal to one-half of the Hamming distance between their *incidence matrices*, namely:

$$\widetilde{\mathscr{P}}(\mathbf{x}, \mathbf{y}) = \widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \triangleq \sum_{1 \le i < l \le n} \left| \delta_{il}(\tilde{\mathbf{x}}) - \delta_{il}(\tilde{\mathbf{y}}) \right|,$$

$$\delta_{il}(\tilde{\mathbf{x}}) = \delta_{il}(\mathbf{x}) \triangleq \begin{cases} 1, & \text{if } x_i = x_l, \\ 0, & \text{if } x_i \ne x_l. \end{cases}$$

Obviously, $\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ satisfies the triangle inequality, i.e., $\widetilde{\mathscr{P}}$-distance is a metric in the space of unordered q-partitions. The following calculation formula for $\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ was established in [2].

Proposition 3 [2]. *For any q-partitions* $\tilde{\mathbf{x}}$ *and* $\tilde{\mathbf{y}}$,

$$\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \frac{1}{2}\sum_{x=0}^{q-1}|E_x(\mathbf{x})|^2 + \frac{1}{2}\sum_{y=0}^{q-1}|E_y(\mathbf{y})|^2 - \sum_{x=0}^{q-1}\sum_{y=0}^{q-1}|E_x(\mathbf{x}) \cap E_y(\mathbf{y})|^2.$$

Let A_q^n, $|A_q^n| = q^n$ *be the set of q-ary n-sequences and we consider* $\widetilde{\mathscr{P}}$*-distance* $\widetilde{\mathscr{P}}(\mathbf{x}, \mathbf{y})$ *as a function of arguments* $\mathbf{x} \in A_q^n$ *and* $\mathbf{y} \in A_q^n$. *This function is a particular case of the concept called U-statistic* (*see, for instance,* [3]). *One can easily prove Proposition* 4.

Proposition 4. *The average value of* $\widetilde{\mathscr{P}}$*-distance is*

$$\overline{\widetilde{\mathscr{P}}(\mathbf{x}, \mathbf{y})} \triangleq \frac{1}{q^{2n}}\sum_{\mathbf{x}\in A_q^n}\sum_{\mathbf{y}\in A_q^n}\widetilde{\mathscr{P}}(\mathbf{x}, \mathbf{y})$$

$$= \frac{n(n-1)}{2}\overline{|\delta_{12}(\mathbf{x}) - \delta_{12}(\mathbf{y})|} = n(n-1)\frac{1}{q}\left(1 - \frac{1}{q}\right).$$

In addition, the distance

$$\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \le \frac{n^2}{2}\left(1 - \frac{1}{q}\right),$$

where the sign of equality is achieved, for example, if $n = sq$, *q-partition* $\tilde{\mathbf{x}} = [n]$ *and for q-partition* $\tilde{\mathbf{y}}$, *all sets* $E_y(\mathbf{y})$, $y \in A_q$, *have the same size* $|E_y(\mathbf{y})| = s$, $s = 1, 2, \ldots$.

Problem (Large Deviations)**.** For parameter $d > 0$, define function

$$\underline{R}_q(\widetilde{\mathscr{P}}, d) \triangleq \overline{\lim_{n\to\infty}} -\frac{1}{n}\log_q\left[\frac{\left|\{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \in A_q^n, \mathbf{y} \in A_q^n, \widetilde{\mathscr{P}}(\mathbf{x}, \mathbf{y}) \le dn^2\}\right|}{q^{2n}}\right].$$

Prove that for any d, $0 < d < \frac{1}{q}(1 - \frac{1}{q})$, the function $\underline{R}_q(\widetilde{\mathscr{P}}, d) > 0$.

Let $h_q(d) \triangleq -d\log_q d - (1-d)\log_q(1-d)$, $0 < d < 1$, be the binary entropy. If $q = 2$, then the function $\underline{R}_2(\widetilde{\mathscr{P}}, d)$, $0 < d < \frac{1}{4}$, is calculated in

Proposition 5 [5]. *For the particular case* $q = 2$, *partition distance* $\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, $0 \le \widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \le n^2/4$, *between 2-partitions* $\tilde{\mathbf{x}}$ *and* $\tilde{\mathbf{y}}$ *is connected with Hamming distance* $H(\mathbf{x}, \mathbf{y})$, $0 \le H(\mathbf{x}, \mathbf{y}) \le n$, *by formula*

$$\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = H(\mathbf{x}, \mathbf{y}) \cdot [n - H(\mathbf{x}, \mathbf{y})]$$

$$\Longleftrightarrow d_H \triangleq \frac{H(\mathbf{x}, \mathbf{y})}{n} = \frac{1}{2}\left(1 \pm \sqrt{1 - 4\frac{\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})}{n^2}}\right).$$

Hence, if $0 < d < \frac{1}{4}$, *then* $\underline{R}_2(\widetilde{\mathscr{P}}, d) > 0$ *and the equality*

$$\underline{R}_2(\widetilde{\mathscr{P}}, d) = 1 - h_2\left(\frac{1-\sqrt{1-4d}}{2}\right), \quad 0 < d < \frac{1}{4},$$

holds.

2. CODES FOR PARTITION DISTANCE

We will say that q-ary $(n \times N)$-matrix $X = \|x_i(j)\|$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, N$, $x_i(j) \in A_q$, is a q-ary *code of length n and size N*. Columns $\mathbf{x}(j) \triangleq (x_1(j), x_2(j), \ldots, x_n(j))$, $j = 1, 2, \ldots, N$, of X are called *codewords* and we will write $X \triangleq \{\mathbf{x}(j), j = 1, 2, \ldots, N\}$. In what follows, we will also interpret code X as a collection of the corresponding q-partitions, i.e., we will also write $X \triangleq \{\tilde{\mathbf{x}}(j), j = 1, 2, \ldots, N\}$, where

$$\tilde{\mathbf{x}}(j) \triangleq \{E_0(\mathbf{x}(j)), E_1(\mathbf{x}(j)), \ldots, E_{q-1}(\mathbf{x}(j))\},$$
$$E_x(\mathbf{x}(j)) \triangleq \{i : x_i(j) = x\}, \quad x \in A_q.$$

Using an analogy with error-correcting codes [4], we give

Definition 3. X is called an *q-partition code for the set* $[n]$ and $\mathscr{P}(X) \triangleq \min_{j \neq j'} \mathscr{P}(\mathbf{x}(j), \mathbf{x}(j'))$ $(\widetilde{\mathscr{P}}(X) \triangleq \min_{j \neq j'} \widetilde{\mathscr{P}}(\mathbf{x}(j), \mathbf{x}(j')))$ is called a $\mathscr{P}(\widetilde{\mathscr{P}})$-*distance* of X.

2.1. Construction

Let q be a prime power and the q-ary alphabet A_q be interpreted as the field F_q with addition $(\oplus)$ and multiplication $(\cdot)$. For vectors $\mathbf{a} \triangleq (a_1, a_2, \ldots, a_m) \in F_q^m$, and $\mathbf{z} \triangleq (z_1, z_2, \ldots, z_m) \in F_q^m$, introduce

$$(\mathbf{a}, \mathbf{z}) \triangleq \bigoplus_{i=1}^{m} (a_i \cdot z_i) \in F_q \quad \text{and} \quad \langle (\mathbf{a}, \mathbf{z}), \mathbf{z} \in F_q^m \rangle \in F_q^n,$$
$$n = q^m, \quad m = 2, 3, \ldots.$$

The first symbol is the standard notation for the *dot product* and the second symbol denotes q-ary sequence of length $n = q^m$ elements of which are dot products of a fixed vector $\mathbf{a} \in F_q^m$ times vectors $\mathbf{z} \in F_q^m$ numbered in the alphabet order.

In this section, we present a construction of q-partition codes X based on the first order Reed–Muller code [4].

Proposition 6. *For* $m = 2, 3, \ldots$, *there exists a family of q-ary codes* $X = X_m$ *of length* $n = q^m$, *size* $N = \frac{q^m - 1}{q-1} + 1$ *and* $\mathscr{P}(\widetilde{\mathscr{P}})$-*distance*

of X_m *is*

$$\mathscr{P}(X_m) = (q-1)q^{m-1}$$
$$= n \cdot \frac{q-1}{q}\left(\widetilde{\mathscr{P}}(X_m) = (q-1)q^{2(m-1)} = n^2 \cdot \frac{1}{q} \cdot \left(1 - \frac{1}{q}\right)\right).$$

Proof. Given a fixed integer $m = 2, 3, \ldots$, denote by $\mathscr{A}$ the maximal set of vectors $\mathbf{a} \in F_q^m$ such that for any $b \in F_q$, $b \neq 0$, and any $\mathbf{a}_1, \mathbf{a}_2 \in \mathscr{A}$, where $\mathbf{a}_1 \neq \mathbf{a}_2$, the inequality $\mathbf{a}_1 \neq b \cdot \mathbf{a}_2$ holds. It is clear that the size of $\mathscr{A}$ is $|\mathscr{A}| = \frac{q^m-1}{q-1} + 1$. Consider the q-ary code[1]

$$X_m \triangleq \{\langle(\mathbf{a}, \mathbf{z}), \mathbf{z} \in F_q^m\rangle \mid \mathbf{a} \in \mathscr{A}\} = \{\mathbf{x}_m(1), \mathbf{x}_m(2), \ldots, \mathbf{x}_m(N)\},$$
$$\mathbf{x}_m(j) = (x_1^{(m)}(j), x_2^{(m)}(j), \ldots, x_n^{(m)}(j)), \quad j = 1, 2, \ldots, N,$$

of length $n = q^m$ and size $N \triangleq |X_m| = |\mathscr{A}| = \frac{q^m-1}{q-1} + 1$. Let codeword $\mathbf{x}_m(1) = (0, 0, \ldots, 0)$ correspond to $\mathbf{a} = \mathbf{0}$, i.e., q-partition $\tilde{\mathbf{x}}_m(1) = [n]$ and codewords $\mathbf{x}_m(j)$, $j = 2, 3, \ldots, N$, identifying q-partitions

$$\tilde{\mathbf{x}}_m(j) = \{E_0(\mathbf{x}_m(j)), E_1(\mathbf{x}_m(j)), \ldots, E_{q-1}(\mathbf{x}_m(j))\},$$
$$E_x(\mathbf{x}_m(j)) \triangleq \{i : x_i^{(m)}(j) = x\}, \quad x \in F_q,$$

correspond to nonzero elements $\mathbf{a} \in \mathscr{A}$, $\mathbf{a} \neq \mathbf{0}$.

Let $\mathbf{a}_1 \neq \mathbf{a}_2$, where $\mathbf{a}_1 \neq \mathbf{0}$ and $\mathbf{a}_2 \neq \mathbf{0}$, are two arbitrary fixed elements of $\mathscr{A}$ and

$$\mathbf{x}_m(j) = \langle(\mathbf{a}_1, \mathbf{z}), \mathbf{z} \in F_q^m\rangle, \qquad \mathbf{x}_m(j') = \langle(\mathbf{a}_2, \mathbf{z}), \mathbf{z} \in F_q^m\rangle$$
$$j, j' = 2, 3, \ldots, N, \quad j \neq j',$$

is the corresponding distinct nonzero codewords of X_m. For any $x \in F_q$, the size of the set $E_x(\mathbf{x}_m(j))$ is:

1) $$|E_x(\mathbf{x}_m(j))| = |\{\mathbf{z} \in F_q^m : (\mathbf{a}_1, \mathbf{z}) = x\}| = q^{m-1}.$$

The last equality follows because $\mathbf{a}_1 \neq \mathbf{0}$. For any $x, y \in F_q$, the size of the intersection $E_x(\mathbf{x}_m(j)) \cap E_y(\mathbf{x}_m(j'))$ is

2) $$|E_x(\mathbf{x}_m(j)) \cap E_y(\mathbf{x}_m(j'))|$$
$$= |\{\mathbf{z} \in F_q^m : (\mathbf{a}_1, \mathbf{z}) = x, (\mathbf{a}_2, \mathbf{z}) = y\}| = q^{m-2}.$$

The last equality follows because system $\{(\mathbf{a}_1, \mathbf{z}) = x, \ (\mathbf{a}_2, \mathbf{z}) = y\}$ of two linear equations over F_q is non-degenerate since $\mathbf{a}_1 \neq b \cdot \mathbf{a}_2$, $\mathbf{a}_1 \neq \mathbf{0}$ and $\mathbf{a}_2 \neq \mathbf{0}$. Hence, the number of solutions to this system is equal to q^{m-2}.

[1]Code X_m is a subcode of the first order Reed–Muller code C_m [4], where C_m is an q-ary linear (n, k)-code, $k = m + 1$, of length $n = q^m$, size $|C_m| = q^{m+1}$ and Hamming distance $d = (q-1) \cdot q^{m-1}$.

Obviously, for any $\mu \in \mathscr{M}_q$, the Hamming distance

3) $H(\mathbf{x}_m(j), \mathbf{x}_m(j')^\mu) = q^m - S(\mathbf{x}_m(j), \mathbf{x}_m(j')^\mu)$,
$j, j' = 1, 2, \ldots, N, \ j \neq j'$,

where $S(\mathbf{x}, \mathbf{y})$ is the Hamming similarity, i.e., the number of coordinates where $\mathbf{x}$ and $\mathbf{y}$ coincide. For codeword $\mathbf{x}_m(1) = (0, 0, \ldots, 0)$ and any codeword $\mathbf{x}_m(j)$, $j = 2, 3, \ldots, N$, the similarity

4) $S(\mathbf{x}_m(1), \mathbf{x}_m(j)^\mu) = |\{\mathbf{z} \in F_q^m : (\mathbf{a}_1, \mathbf{z}) = \mu^{-1}(0)\}| \overset{1)}{=} q^{m-1}$.

If $j, j' = 2, 3, \ldots, N$, $j \neq j'$, then

5) $S(\mathbf{x}_m(j), \mathbf{x}_m(j')^\mu)$
$= \sum_{x \in F_q} |\{\mathbf{z} \in F_q^m : (\mathbf{a}_1, \mathbf{z}) = x, (\mathbf{a}_2, \mathbf{z}) = \mu^{-1}(x)\}|$
$\overset{2)}{=} q \cdot q^{m-2} = q^{m-1}$.

From 3)–5) it follows that, for any $\mu \in \mathscr{M}_q$ and any two codewords $\mathbf{x}_m(j) \neq \mathbf{x}_m(j')$ of code X_m, the Hamming distance

6) $H(\mathbf{x}_m(j), \mathbf{x}_m(j')^\mu) = q^m - S(\mathbf{x}_m(j), \mathbf{x}_m(j')^\mu) = q^m - q^{m-1}$
$= (q-1) \cdot q^{m-1}$.

If $\tilde{\mathbf{x}}_m(1) = [n] = [q^m]$, then in virtue of 1) the formula of Proposition 3 has the form

7. $\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}_m(1), \tilde{\mathbf{x}}_m(j)) = \frac{1}{2} \cdot q^{2m} + \frac{1}{2} q \cdot q^{2m-2} - q \cdot q^{2m-2} = \frac{q-1}{2} q^{2m-1}$.

If $j, j' = 2, 3, \ldots, N$, $j \neq j'$, then in virtue of 1)–2) the formula of Proposition 3 yields

8) $\widetilde{\mathscr{P}}(\tilde{\mathbf{x}}_m(j), \tilde{\mathbf{x}}_m(j')) = \frac{1}{2} q \cdot q^{2m-2} + \frac{1}{2} q \cdot q^{2m-2} - q^2 \cdot q^{2m-4}$
$= (q-1) \cdot q^{2(m-1)}$.

Proposition 6 follows from 6)–8).

Remark 2. Propositions 1 and 6 mean that code X_m has the maximal possible $\mathscr{P}$-distance. In virtue of formula 6), code X_m is an equidistant code in $\mathscr{P}$-metrics. From Propositions 4 and 6 it follows that $\widetilde{\mathscr{P}}$-distance of X_m asymptotically ($n \to \infty$) coincides with the average $\widetilde{\mathscr{P}}$-distance.

2.2. Random Coding Bound for $\mathscr{P}$-Distance

Let D, $1 \leq D \leq n(1 - 1/q)$, be an integer and $N_q(\mathscr{P}, n, D)$ denote the *maximal size* of q-partition codes X for the set $[n]$ having distance $\mathscr{P}(X) \geq D$.

If d, $0 < d < 1 - 1/q$, is fixed, then

$$R_q(\mathscr{P}, d) \triangleq \overline{\lim_{n\to\infty}} \frac{\log_q N_q(\mathscr{P}, n, dn)}{n}$$

is called a *rate* of q-partition codes for $\mathscr{P}$-distance. A random coding lower bound on $R_q(\mathscr{P}, d)$ presented by Proposition 7 coincides with the classical Varshamov–Gilbert bound [4] for codes in Hamming metric.

Proposition 7. *For any d, $0 < d < 1 - 1/q$, the rate $R_q(\mathscr{P}, d) > 0$ and*

$$R_q(\mathscr{P}, d) \geq \underline{R}_q(\mathscr{P}, d) \triangleq 1 - d\log_q(q-1) - h_q(d),$$
$$h_q(d) \triangleq -d\log_q d - (1-d)\log_q(1-d).$$

Proof. Let $\mathbf{x}(j)$ and $\mathbf{x}(j')$, $j \neq j'$, be q-ary independent random codewords of code X with the same uniform distribution, i.e., for any q-ary n-sequence $\mathbf{x}$, the probability

$$\Pr\{\mathbf{x}(j) = \mathbf{x}\} = \Pr\{\mathbf{x}(j') = \mathbf{x}\} = q^{-n}.$$

For $\mu \in \mathscr{M}_q$, consider the random variable $\xi_\mu = \xi_\mu(\mathbf{x}(j), \mathbf{x}(j')) \triangleq \sum_{x\in A_q} n(x, \mu(x) \mid \mathbf{x}(j), \mathbf{x}(j'))$. One can easily check that ξ_μ has the following binomial distribution:

$$\Pr\{\xi_\mu(\mathbf{x}(j), \mathbf{x}(j')) = k\} = \binom{n}{k} \cdot \left(\frac{1}{q}\right)^k \cdot \left(1 - \frac{1}{q}\right)^{n-k}, \quad k = 0, 1, \ldots, n,$$

which does not depend on μ. In virtue of Definition 1 and Proposition 1, the partition distance $\mathscr{P}(\mathbf{x}(j), \mathbf{x}(j')) = n - \max_{\mu\in\mathscr{M}_q} \xi_\mu(\mathbf{x}(j), \mathbf{x}(j'))$ and, therefore, for any $\mu \in \mathscr{M}_q$, the probability

$$\Pr\{\mathscr{P}(\mathbf{x}(j), \mathbf{x}(j')) \leq nd\} \leq q! \sum_{k=n(1-d)}^{n} \binom{n}{k} \cdot \left(\frac{1}{q}\right)^k \cdot \left(1 - \frac{1}{q}\right)^{n-k}.$$

If d, $0 < d < 1 - 1/q$, is fixed, then the standard arguments used to obtain the random coding bound yield

$$\begin{aligned} R_q(\mathscr{P}, d) \geq \underline{R}_q(\mathscr{P}, d) &\triangleq \overline{\lim_{n\to\infty}} - \frac{\log_q \left[\sum_{k=n(1-d)}^{n} \binom{n}{k} \cdot \left(\frac{1}{q}\right)^k \cdot \left(1 - \frac{1}{q}\right)^{n-k}\right]}{n} \\ &= -d\log_q\left(1 - \frac{1}{q}\right) - (1-d)\log_q\frac{1}{q} - h_q(d) \\ &= 1 - d\log_q(q-1) - h_q(d). \end{aligned}$$

Proposition 7 is proved.

2.3. Bounds for $\widetilde{\mathscr{P}}$-Distance

Let D, $1 \le D \le \frac{n^2}{2}(1-\frac{1}{q})$ be an integer and $N_q(\widetilde{\mathscr{P}}, n, D)$ denote the *maximal size* of q-partition codes X for the set $[n]$ having $\widetilde{\mathscr{P}}$-distance $\widetilde{\mathscr{P}}(X) \ge D$. If d, $0 < d < \frac{1}{2}(1-\frac{1}{q})$, is fixed, then

$$R_q(\widetilde{\mathscr{P}}, d) \triangleq \overline{\lim_{n\to\infty}} \frac{\log_q N_q(\widetilde{\mathscr{P}}, n, dn^2)}{n}$$

is called a *rate* of q-partition codes for $\widetilde{\mathscr{P}}$-distance.

In general case $q \ge 2$, a random coding lower bound on the rate of q-partition codes for $\widetilde{\mathscr{P}}$-distance has the form

$$R_q(\widetilde{\mathscr{P}}, d) \ge \underline{R}_q(\widetilde{\mathscr{P}}, d), \quad 0 < d < \frac{1}{q}\left(1-\frac{1}{q}\right)$$

where the unknown function $\underline{R}_q(\widetilde{\mathscr{P}}, d)$ was defined in Section 1.2 for the corresponding problem of large deviations in the theory of U-statistics. The given bound is an analog of the Varshamov–Gilbert lower bound for Hamming metric.

For the particular case $q = 2$ the function $\underline{R}_2(\widetilde{\mathscr{P}}, d) = 1 - h_2(\frac{1-\sqrt{1-4d}}{2})$ was calculated in Proposition 5. This yields

Proposition 8 (Varshamov–Gilbert Lower Bound for $\widetilde{\mathscr{P}}$-Distance)**.**

$$R_2(\widetilde{\mathscr{P}}, d) \ge 1 - h_2\left(\frac{1-\sqrt{1-4d}}{2}\right), \quad 0 < d < \frac{1}{4}.$$

Let $q = 2$ and $R_2(d_H)$, $0 < d_H < 1/2$ be the rate of conventional binary Hamming metric codes. The following upper bound called Elias bound is known [4]:

$$R_2(d_H) \le 1 - h_2\left(\frac{1-\sqrt{1-2d_H}}{2}\right), \quad 0 < d_H < \frac{1}{2}.$$

Therefore, in virtue of Proposition 5, we obtain

Proposition 9 (Elias Upper Bound for $\widetilde{\mathscr{P}}$-Distance)**.**

$$R_2(\widetilde{\mathscr{P}}, d) \le 1 - h_2\left(\frac{1-\sqrt[4]{1-4d}}{2}\right), \quad 0 < d < \frac{1}{4}.$$

3. APPLICATIONS

Given integers $2 \le q < n$, consider an arbitrary q-partition code $X = \{\tilde{\mathbf{x}}(j), j = 1, 2 \ldots, N\}$ for the set $[n]$. By symbol $\mathscr{D} = \mathscr{D}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$, we will

denote partition distance $\mathscr{P} = \mathscr{P}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ or partition distance $\widetilde{\mathscr{P}} = \widetilde{\mathscr{P}}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ introduced in Section 1 (see Definitions 1 and 2). Let

$$D \triangleq \max_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})} \mathscr{D}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \leq \begin{cases} \lfloor (q-1)n/q \rfloor, & \text{if } \mathscr{D} = \mathscr{P}, \\ \lfloor (q-1)n^2/2q \rfloor, & \text{if } \mathscr{D} = \widetilde{\mathscr{P}}. \end{cases}$$

be the maximal $\mathscr{D}$-distance between q-partitions of the set $[n]$.

3.1. Statistical analysis of q-partition samples

The following psychological or medical testing can be called a "nonparametric" (n, q)-questionnaire.[2]

- An individual or patient is asked to split n objects up into q' groups, $1 \leq q' \leq q$, putting two seemingly similar objects in the same group.
- We test M individuals and get a *sample* $Y = \{\tilde{\mathbf{y}}(m), m = 1, 2, \ldots M\}$, containing M observations, i.e., q-partitions of the set $[n]$. Next, we apply a statistical analysis of q-partition samples based on $\mathscr{D}$-distance.
- Let r, $r = 0, 1, \ldots, D$ be an integer and $\tilde{\mathbf{x}}$ be an arbitrary q-partition. Introduce sets:

$$\widetilde{S}_n^q(r, \tilde{\mathbf{x}}) \triangleq \{\tilde{\mathbf{y}} : \mathscr{D}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = r\}, \quad \widetilde{B}_n^q(r, \tilde{\mathbf{x}}) \triangleq \sum_{i=0}^{r} \widetilde{S}_n^q(i, \tilde{\mathbf{x}}),$$

$$S_n^q(r, \tilde{\mathbf{x}}) \triangleq \{\mathbf{y} : \mathscr{D}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = r\}, \quad B_n^q(r, \tilde{\mathbf{x}}) \triangleq \sum_{i=0}^{r} S_n^q(i, \tilde{\mathbf{x}}).$$

  The first two sets are a *sphere* and a *ball* with *center* $\tilde{\mathbf{x}}$ and *radius* r. The second two sets are sets of q-ary n-sequences identifying q-partitions of the given sphere and ball. Obviously, if $r = D$, then the size $|B_n^q(D, \tilde{\mathbf{x}})| = q^n$ and if $\tilde{\mathbf{x}} = [n]$, then $|S_n^q(0, [n])| = q$ and $|\widetilde{S}_n^q(0, [n])| = 1$.
- Consider the sample $Y = \{\tilde{\mathbf{y}}(m), m = 1, 2, \ldots M\}$. Fix an arbitrary q-partition $\tilde{\mathbf{x}}$, and integer $r = 1, 2, \ldots, D$. Let $m(r) = m(r, \tilde{\mathbf{x}}, Y)$ denote the number of indices m, $m = 1, 2, \ldots, M$, for which $\tilde{\mathbf{y}}(m) \in \widetilde{B}_n^q(r, \tilde{\mathbf{x}})$, i.e., $\mathscr{D}$-distance $\mathscr{D}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}(m)) \leq r$. Therefore, we have *statistics*: $0 \leq m(1) \leq m(2) \leq \cdots \leq m(D) = M$.

[2]The conventional "parametric" (n, q)-questionnaire contains n questions and an individual (patient) gives one of q possible answers to each question. Each answer is estimated by the corresponding number $0, 1, \ldots, q-1$ and the patient gets a tentative diagnosis identified by the sum (score) of n estimates. One has to take into account that this conventional approach cannot be applied to nonparametric questionnaires.

- For an integer $r = 1, 2, \ldots, D-1$, denote by p_r an unknown probability of a success in M independent Bernoulli trials where the number of successes is $m(r)$. If q-partitions $\tilde{\mathbf{y}}(m)$ are identified by independent q-ary n-sequences $\mathbf{y}(m)$, $m = 1, 2, \ldots M$, having the uniform distribution (hypothesis $\mathbf{H_0}$) in the space of all q-ary n-sequences, then the unknown probability

$$p_r = u_r = u_r(\tilde{\mathbf{x}}) \triangleq \frac{|B_n^q(r, \tilde{\mathbf{x}})|}{q^n} = \sum_{i=0}^{r} |S_n^q(i, \tilde{\mathbf{x}})|/q^n, \quad 0 < u_r < 1.$$

- On the base of statistic $m(r)$ we test the *homogeneous* hypothesis

$$\mathbf{H}_0 : p_r = u_r \quad \text{versus} \quad \mathbf{H}_1 : p_r > u_r (p_r < u_r).$$

We say: $\mathbf{H}_1$ *means that q-partition* $\tilde{\mathbf{x}}$ *is a center of r-concentration* (*r-vacuum*) *for sample Y and hypothesis* $\mathbf{H}_1$ *is accepted at the level of significance* $\ell_c(r) = \ell_c(r, \tilde{\mathbf{x}}, Y)(\%)(\ell_v(r) = \ell_v(r, \tilde{\mathbf{x}}, Y)(\%))$, *where*

$$\frac{\ell_c(r)}{100} \triangleq \sum_{i=m(r)}^{M} \binom{M}{i} \cdot u_r^i \cdot (1-u_r)^{M-i}$$

$$\left(\frac{\ell_v(r)}{100} \triangleq \sum_{i=0}^{m(r)} \binom{M}{i} \cdot u_r^i \cdot (1-u_r)^{M-i}\right).$$

- We suggest to find centers of r-concentration and r-vacuum for sample Y among codewords of an q-partition code $X = \{\tilde{\mathbf{x}}(j), j = 1, 2 \ldots, N\}$ for the set $[n]$ having distance $\mathscr{D}(X)$, which is close to its maximum $\mathscr{D}(n; N)$. For instance, if q is a prime power, then the bound and construction obtained in Section 2.1 yield the maximum of $\mathscr{P}$-distance in the form:

$$\mathscr{P}\left(q^m; \frac{q^m - 1}{q-1} + 1\right) = q^{m-1}(q-1), \quad m = 2, 3, \ldots.$$

- Given a number $\ell(\%)$ and an integer $r = 1, 2, \ldots, D-1$, define subcode $X_r(\ell, Y) \subseteq X$ as follows:

$$X_r(\ell, Y) \triangleq \{\tilde{\mathbf{x}} : \tilde{\mathbf{x}} \in X \text{ and } \min[\ell_c(r, \tilde{\mathbf{x}}, Y); \ell_v(r, \tilde{\mathbf{x}}, Y)] \leq \ell\},$$

i.e., subcode $X_r(\ell, Y)$ contains all centers of r-concentration and r-vacuum for sample Y at the level of significance $\leq \ell(\%)$.

- This gives a possibility to *separate* samples Y on the base of a *comparison* of the corresponding subcodes $X_r(\ell, Y)$ for the standard levels of significance $\ell \leq 5\%$.

3.2. Example

Let $q = 3$, $n = 9$, and 3-partition code $X = \{\tilde{\mathbf{x}}(j), j = 1, 2 \ldots, 5\}$ for the set [9] have the form:

$$\tilde{\mathbf{x}}(1) = \{(1, 2, 3, 4, 5, 6, 7, 8, 9)\}, \quad \tilde{\mathbf{x}}(2) = \{(1, 2, 3); (4, 5, 6); (7, 8, 9)\},$$
$$\tilde{\mathbf{x}}(3) = \{(1, 4, 7); (2, 5, 8); (3, 6, 9)\}, \quad \tilde{\mathbf{x}}(4) = \{(1, 6, 8); (2, 4, 9); (3, 5, 7)\},$$
$$\tilde{\mathbf{x}}(5) = \{(1, 5, 9); (2, 6, 7); (3, 4, 8)\}.$$

One can easily check that this code is a particular case of the construction presented in Proposition 6 from Section 2. Hence, $\mathscr{P}(\widetilde{\mathscr{P}})$-distance of code X is

$$\mathscr{P}(X) = \mathscr{P}(\mathbf{x}(j), \mathbf{x}(j')) = 6, \quad 1 \le j < j' \le 5,$$
$$(\widetilde{\mathscr{P}}(X) = \widetilde{\mathscr{P}}(\mathbf{x}(j), \mathbf{x}(j')) = 18, \quad 2 \le j < j' \le 5).$$

In addition, for the metric $\mathscr{P}$ $(\widetilde{\mathscr{P}})$, the maximal distance D between q-partitions of the set $[n]$ is

$$D = \mathscr{P}(X) = 6 \quad (D = \widetilde{\mathscr{P}}(\mathbf{x}(1), \mathbf{x}(j)) = 27, \ 2 \le j \le 5).$$

We will apply the code $X = \{\tilde{\mathbf{x}}(j), j = 1, 2, \ldots, 5\}$ to the statistical analysis of sample $Y = \{\tilde{\mathbf{y}}(m), m = 1, 2, \ldots, 27\}$ containing the following 3-partitions of the set $[9] = \{1, 2, \ldots, 9\}$:

$$\tilde{\mathbf{y}}(1) = \{(1); (2, 3, 4, 5, 6, 7, 8); (9)\},$$
$$\tilde{\mathbf{y}}(2) = \{(1, 3, 4); (2, 5, 7); (6, 8, 9)\},$$
$$\tilde{\mathbf{y}}(3) = \{(1, 2, 3, 4, 5, 6, 7, 9); (8)\},$$
$$\tilde{\mathbf{y}}(4) = \{(1, 2, 3, 6, 7, 9); (4, 5); (8)\},$$
$$\tilde{\mathbf{y}}(5) = \{(1, 3, 7); (2, 9); (4, 5, 6, 8)\},$$
$$\tilde{\mathbf{y}}(6) = \{(1, 2, 3, 4, 5, 6, 7, 9); (8)\},$$
$$\tilde{\mathbf{y}}(7) = \{(1, 3, 4, 5, 7, 8); (2, 6, 9)\},$$
$$\tilde{\mathbf{y}}(8) = \{(1, 2, 3, 6, 8); (4, 5, 7, 9)\},$$
$$\tilde{\mathbf{y}}(9) = \{(1, 7); (2, 8); (3, 4, 5, 6, 9)\},$$
$$\tilde{\mathbf{y}}(10) = \{(1); (2, 4, 5, 7, 8, 9); (3, 6)\},$$
$$\tilde{\mathbf{y}}(11) = \{(1); (2, 3, 4, 5, 7, 8); (6, 9)\},$$
$$\tilde{\mathbf{y}}(12) = \{(1, 3); (2, 4, 5, 6, 7); (8, 9)\},$$
$$\tilde{\mathbf{y}}(13) = \{(1, 3, 6, 7, 8); (2, 4); (5, 9)\},$$
$$\tilde{\mathbf{y}}(14) = \{(1, 7); (2, 3, 4, 5, 6); (8, 9)\},$$

$$\tilde{\mathbf{y}}(15) = \{(1, 9); (2, 3, 4, 5, 6, 7, 8)\},$$
$$\tilde{\mathbf{y}}(16) = \{(1, 6, 7, 8); (2, 3, 4, 5, 9)\},$$
$$\tilde{\mathbf{y}}(17) = \{(1, 2, 3, 5, 6, 7, 8, 9); (4)\},$$
$$\tilde{\mathbf{y}}(18) = \{(1, 3, 6, 7); (2, 4, 5, 8); (9)\},$$
$$\tilde{\mathbf{y}}(19) = \{(1, 7); (2, 4, 5, 6); (3, 8, 9)\},$$
$$\tilde{\mathbf{y}}(20) = \{(1, 4, 5); (2, 3, 6, 7, 8, 9)\},$$
$$\tilde{\mathbf{y}}(21) = \{(1, 2, 3, 4, 8, 9); (5, 6, 7)\},$$
$$\tilde{\mathbf{y}}(22) = \{(1, 2, 4, 5, 7, 9); (3); (6, 8)\},$$
$$\tilde{\mathbf{y}}(23) = \{(1, 2, 5, 7, 8, 9); (3, 4, 6)\},$$
$$\tilde{\mathbf{y}}(24) = \{(1, 3, 7); (2, 4, 6, 8); (5, 9)\},$$
$$\tilde{\mathbf{y}}(25) = \{(1, 2, 4, 5, 8, 9); (3); (6, 7)\},$$
$$\tilde{\mathbf{y}}(26) = \{(1); (2, 3, 4, 5, 7, 9); (6, 8)\},$$
$$\tilde{\mathbf{y}}(27) = \{(1, 3, 6, 7); (2, 4, 8); (5, 9)\}.$$

The sample was taken from a protocol of the psychological tests with $M = 27$ individuals for which the estimates of $n = 9$ *nature accentuation* types were calculated. The pairs of distances $[\mathscr{P}(\tilde{\mathbf{x}}(i), \tilde{\mathbf{y}}(j))\ \widetilde{\mathscr{P}}(\tilde{\mathbf{x}}(i), \tilde{\mathbf{y}}(j))]$, $1 \le i \le 5$, $1 \le j \le 27$, are given in Table 1, where we marked in boldface type minimal distances identifying clusters of sample Y generated by code X.

Table 1. Pairs of distances $[\mathscr{P}(\tilde{\mathbf{x}}(i), \tilde{\mathbf{y}}(j))\ \widetilde{\mathscr{P}}(\tilde{\mathbf{x}}(i), \tilde{\mathbf{y}}(j))]$

	$\tilde{\mathbf{x}}(1)$		$\tilde{\mathbf{x}}(2)$		$\tilde{\mathbf{x}}(3)$		$\tilde{\mathbf{x}}(4)$		$\tilde{\mathbf{x}}(5)$	
$\tilde{\mathbf{y}}(1)$	**2**	**15**	4	20	4	20	4	20	5	18
$\tilde{\mathbf{y}}(2)$	6	27	4	14	**3**	**12**	4	14	4	19
$\tilde{\mathbf{y}}(3)$	**1**	**8**	5	23	5	23	5	23	5	23
$\tilde{\mathbf{y}}(4)$	3	20	3	**15**	4	17	5	19	4	17
$\tilde{\mathbf{y}}(5)$	5	26	3	**11**	4	15	3	13	5	17
$\tilde{\mathbf{y}}(6)$	**1**	**8**	5	23	5	23	5	23	5	23
$\tilde{\mathbf{y}}(7)$	**3**	18	6	21	4	17	4	17	4	17
$\tilde{\mathbf{y}}(8)$	4	20	4	15	5	19	4	15	5	19
$\tilde{\mathbf{y}}(9)$	4	24	4	15	**2**	**11**	5	17	5	17
$\tilde{\mathbf{y}}(10)$	3	20	4	17	3	**15**	4	17	5	24
$\tilde{\mathbf{y}}(11)$	3	20	5	19	3	**15**	4	17	4	17
$\tilde{\mathbf{y}}(12)$	4	24	**2**	**11**	5	17	5	17	4	15
$\tilde{\mathbf{y}}(13)$	4	24	5	17	5	17	**3**	**11**	4	15
$\tilde{\mathbf{y}}(14)$	4	24	**3**	**11**	4	15	5	17	5	17

(continued)

Table 1. Continued

	$\tilde{\mathbf{x}}(1)$		$\tilde{\mathbf{x}}(2)$		$\tilde{\mathbf{x}}(3)$		$\tilde{\mathbf{x}}(4)$		$\tilde{\mathbf{x}}(5)$	
$\tilde{\mathbf{y}}(15)$	**2**	**14**	5	21	5	21	5	21	4	17
$\tilde{\mathbf{y}}(16)$	4	20	5	19	5	19	**3**	**11**	5	19
$\tilde{\mathbf{y}}(17)$	**1**	**8**	5	23	5	23	5	23	5	23
$\tilde{\mathbf{y}}(18)$	5	24	4	17	**3**	**11**	5	15	4	17
$\tilde{\mathbf{y}}(19)$	5	26	3	**11**	3	13	5	17	4	15
$\tilde{\mathbf{y}}(20)$	**3**	18	4	17	4	17	6	21	4	17
$\tilde{\mathbf{y}}(21)$	**3**	18	4	17	6	21	4	17	4	17
$\tilde{\mathbf{y}}(22)$	3	20	5	19	4	17	3	**15**	4	17
$\tilde{\mathbf{y}}(23)$	**3**	18	4	17	4	17	6	21	4	17
$\tilde{\mathbf{y}}(24)$	5	26	4	15	4	15	4	13	4	13
$\tilde{\mathbf{y}}(25)$	3	20	5	19	4	17	4	17	3	**15**
$\tilde{\mathbf{y}}(26)$	**3**	20	5	19	5	19	4	**11**	5	19
$\tilde{\mathbf{y}}(27)$	5	26	5	17	4	13	4	13	**3**	13
–	–		–		–		–		–	

3.2.1. Statistical Analysis for $\mathscr{P}$-Distance

From Table 1, we obtain statistics $m(r) = m(r, \tilde{\mathbf{x}}(j), Y)$, $j = 1, 2, \ldots, 5$, $r = 0, 1, \ldots, 6$. For codeword $\tilde{\mathbf{x}}(1)$, these statistics and sizes of sets $|B_9^3(r, \tilde{\mathbf{x}}(1))|$ yield levels of significance $\ell_c(r, \tilde{\mathbf{x}}(1), Y)$ and $\ell_v(r, \tilde{\mathbf{x}}(1), Y)$, which are calculated in Table 2. For codewords $\tilde{\mathbf{x}}(j)$, $j = 2, 3, 4, 5$, the corresponding levels of significance $\ell_c(r, \tilde{\mathbf{x}}(j), Y)$ and $\ell_v(r, \tilde{\mathbf{x}}(j), Y)$ are presented in Table 3.

Conclusions. For $\ell = 2.4\%$ and $r = 3$, we generate subcode $X_r(\ell, Y) = \{\tilde{\mathbf{x}}(\mathbf{1}), \tilde{\mathbf{x}}(5)\}$. Codeword $\tilde{\mathbf{x}}(1)$ is a center of 3-concentration for sample Y at the level of significance $< 0.5\%$. Codeword $\tilde{\mathbf{x}}(5)$ is a center of 3-vacuum for sample Y at the level of significance 2.4%.

Table 2. Statistics, sizes of sets $|B_9^3(r, \tilde{\mathbf{x}}(1))|$ and levels of significance for codeword $\tilde{\mathbf{x}}(1)$

r	1	2	3	4	5
$\lvert B_9^3(r, \tilde{\mathbf{x}}(1))\rvert$	57	489	2505	8553	18003
$u_r(\tilde{\mathbf{x}}(1))$	.00290	.0248	.127	.435	.915
$m(r, \tilde{\mathbf{x}}(1), Y)$	3	5	15	21	26
$\ell_c(r, \tilde{\mathbf{x}}(1), Y)$	$<$**0.5**%	**0.05**%	$<$**0.5**%	$<$**0.5**%	32%
$\ell_v(r, \tilde{\mathbf{x}}(1), Y)$	$>$99.5%	$>$99.9%	$>$99.5%	$>$99.5%	91%

Table 3. Statistics, sizes of sets $|B_9^3(r, \tilde{\mathbf{x}}(j))|$ and levels of significance for codewords $\tilde{\mathbf{x}}(j)$, $j = 2, 3, 4, 5$

r	2	3	4	5
$\lvert B_9^3(r, \tilde{\mathbf{x}}(j))\rvert$	978	4830	13686	19302
$u_r(\tilde{\mathbf{x}}(j))$	.0497	.245	.695	.981
$m(r, \tilde{\mathbf{x}}(2), Y)$	1	5	15	26
$\ell_c(r, \tilde{\mathbf{x}}(2), Y)$	75%	83%	95.8%	90.4%
$\ell_v(r, \tilde{\mathbf{x}}(2), Y)$	61%	32%	9%	41%
$m(r, \tilde{\mathbf{x}}(3), Y)$	1	6	17	26
$\ell_c(r, \tilde{\mathbf{x}}(3), Y)$	75%	68%	82%	90.4%
$\ell_v(r, \tilde{\mathbf{x}}(3), Y)$	61%	50%	31%	41%
$m(r, \tilde{\mathbf{x}}(4), Y)$	0	4	15	25
$\ell_c(r, \tilde{\mathbf{x}}(4), Y)$	100%	93%	95.8%	98.5%
$\ell_v(r, \tilde{\mathbf{x}}(4), Y)$	25%	17%	9%	9%
$m(r, \tilde{\mathbf{x}}(5), Y)$	0	2	16	27
$\ell_c(r, \tilde{\mathbf{x}}(5), Y)$	100%	99.5%	91%	59%
$\ell_v(r, \tilde{\mathbf{x}}(5), Y)$	25%	**2.4**%	18%	100%

Comments on Table 2. If $r = 1, 2, \ldots, \lfloor 2n/3 \rfloor$, then $|B_n^3(r, [n])| = \sum_{i=0}^{r} |S_n^3(i, [n])|$ are calculated using the following formulas:

$$|S_n^3(r, [n])| = 6 \cdot |\widetilde{S}_n^3(r, [n])|,$$

$$|\widetilde{S}_n^3(r, [n])| = \sum_{i=|2r-n|^+}^{\lfloor r/2 \rfloor} M(i, r-i, n-r), \quad 1 \le r \le \lfloor 2n/3 \rfloor,$$

where $|a|^+ = \max(a, 0)$ and

$$M(n_0, n_1, n_2) = \begin{cases} \binom{n}{n_2} \cdot \binom{n-n_2}{n_0}, & \text{if } n_0 < n_1 < n_2, \\ \binom{n}{n_2} \cdot \frac{1}{2} \cdot \binom{n-n_2}{n_0}, & \text{if } n_0 = n_1 < n_2, \\ \binom{n}{n_0} \cdot \frac{1}{2} \cdot \binom{n-n_0}{n_1}, & \text{if } n_0 < n_1 = n_2, \\ \binom{n-1}{n_0-1} \cdot \frac{1}{2} \cdot \binom{n-n_0}{n_0}, & \text{if } n_0 = n_1 = n_2. \end{cases}$$

Comments on Table 3. For any codeword $\tilde{\mathbf{x}}(j)$, $j = 2, 3, 4, 5$, sizes o sets $|B_9^3(r, \tilde{\mathbf{x}}(j))|$ do not depend on j. They are calculated using exhaustive search.

Table 4. Results of analysis for metric $\mathscr{P}$ for codeword $\tilde{\mathbf{x}}(1)$

r	8	14	15	18	20	24	26
$\lvert B_9^3(r, \tilde{\mathbf{x}}(1))\rvert$	57	273	489	993	3261	10443	18003
$u_r(\tilde{\mathbf{x}}(1))$	.00290	.0139	.0248	.0504	.166	.531	.915
$m(r, \tilde{\mathbf{x}}(1), Y)$	3	4	5	9	17	22	26
$\ell_c(r, \tilde{\mathbf{x}}(1), Y)$	**<0.01%**	**0.05%**	**0.05%**	**<0.01%**	**<0.01%**	**0.2%**	32%
$\ell_v(r, \tilde{\mathbf{x}}(1), Y)$	>99.9%	>99.9%	>99.9%	>99.9%	>99.9%	>99.9%	91%

3.2.2. Statistical Analysis for $\widetilde{\mathscr{P}}$-Distance

Table 1 yields statistics $m(r) = m(r, \tilde{\mathbf{x}}(j), Y)$, $j = 1, 2, \ldots, 5$, $8 \le r \le 26$. The results of analysis for metric $\widetilde{\mathscr{P}}$ are summarized in Tables 4 (for codeword $\tilde{\mathbf{x}}(1)$) and in Table 5 (for codewords $\tilde{\mathbf{x}}(j)$, $j = 2, 3, 4, 5$), where sizes o sets $\lvert B_9^3(r, \tilde{\mathbf{x}}(j))\rvert$, $j = 1, 2, \ldots, 5$, are calculated using exhaustive search.

Conclusions 1. For $\ell = 0.3\%$ and $r = 14$, subcode $X_r(\ell, Y) = \{\tilde{\mathbf{x}}(1), \tilde{\mathbf{x}}(5)\}$. Codeword $\tilde{\mathbf{x}}(1)$ ($\tilde{\mathbf{x}}(5)$) is a center of 14-concentration (14-vacuum) at the level of significance 0.05% (0.3%).

2. For $\ell = 1.4\%$ and $r = 15$, subcode $X_r(\ell, Y) = \{\tilde{\mathbf{x}}(1), \tilde{\mathbf{x}}(2), \tilde{\mathbf{x}}(5)\}$. Codeword $\tilde{\mathbf{x}}(1)$ ($\tilde{\mathbf{x}}(5)$) is a center of 15-concentration (15-vacuum) at

Table 5. Results of analysis for metric $\mathscr{P}$ for codeword $\tilde{\mathbf{x}}(j)$

r	9	11	12	13	14	15	17	18
$\lvert B_9^3(r, \tilde{\mathbf{x}}(j))\rvert$	996	2616	3264	5694	7314	11040	16926	18168
$u_r(\tilde{\mathbf{x}}(j))$	.0506	.133	.166	.289	.372	.561	.860	.923
$m(r, \tilde{\mathbf{x}}(2), Y)$	0	4	4	4	5	9	16	16
$\ell_c(r, \tilde{\mathbf{x}}(2), Y)$	100%	49%	68%	97%	99%	>99.5%	>99.9%	>99.9%
$\ell_v(r, \tilde{\mathbf{x}}(2), Y)$	25%	71%	53%	7.4%	**3.1%**	**1.4%**	**0.1%**	**<0.01%**
$m(r, \tilde{\mathbf{x}}(3), Y)$	0	2	3	5	5	10	18	18
$\ell_c(r, \tilde{\mathbf{x}}(3), Y)$	100%	98%	85%	93%	99%	98.6%	99.8%	>99.9%
$\ell_v(r, \tilde{\mathbf{x}}(3), Y)$	25%	28%	32%	16%	**3.1%**	**3.6%**	**0.9%**	**<0.01%**
$m(r, \tilde{\mathbf{x}}(4), Y)$	0	3	3	6	7	10	19	19
$\ell_c(r, \tilde{\mathbf{x}}(4), Y)$	100%	72%	85%	84%	92%	98.6%	99.1%	>99.9%
$\ell_v(r, \tilde{\mathbf{x}}(4), Y)$	25%	51%	32%	30%	16%	**3.6%**	**2.8%**	**0.1%**
$m(r, \tilde{\mathbf{x}}(5), Y)$	0	0	0	2	3	7	19	20
$\ell_c(r, \tilde{\mathbf{x}}(5), Y)$	100%	100%	100%	99.9%	>99.9%	>99.9%	99.1%	99.9%
$\ell_v(r, \tilde{\mathbf{x}}(5), Y)$	25%	**2.1%**	**0.7%**	**0.7%**	**0.3%**	**0.1%**	**2.8%**	**0.4%**

the level of significance 0.05% (0.1%). Codeword $\tilde{\mathbf{x}}(2)$ is a center of 15-vacuum at the level of significance 1.4%.

REFERENCES

1. Gusfield, D. 2002. Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters* 82:159–164.
2. Mirkin, B.G., and L.B. Tcherny. 1970. On measurement of proximity between various partitions of finite set. *Avtomatika i Telemekhanika* 5:120–127 (in Russian).
3. Koroljuk, V.S., and Y.V. Borovskich. 1993. *The Theory of U-statistics (Mathematics and its Applications)*. Berlin: Springer (in Russian).
4. MacWilliams, F.J., and N.J.A. Sloan. 1977. *The Theory of Error-Correcting Codes*. Amsterdam: North Holland.
5. Ismagilov, I.K. 2004. Private Correspondence.