DOI 10.1007/s10559-015-9774-1
*Cybernetics and Systems Analysis, Vol. 51, No. 5, September, 2015*

# ESTIMATION OF VECTORS SIMILARITY BY THEIR RANDOMIZED BINARY PROJECTIONS

**D. A. Rachkovskij**

UDC 004.22 + 004.93'11

**Abstract.** *We analyze the estimation of the angle, scalar product, and the Euclidean distance of real-valued vectors using binary vectors with controlled sparsity. Transformation is carried out by projection using a binary random matrix with elements* $\{0,1\}$ *and the output threshold transformation. We also provide a comparative analysis of the error obtained while estimating the similarity measures of input vectors by some similarity measures of output binary vectors based on their scalar product.*

**Keywords:** *binary random projections, sparse binary representations, estimate of vector similarity.*

## INTRODUCTION

Vectorial data representation seems to be the most popular in information technologies. Many search methods using similarity, classification, regression, clustering, as well as methods of estimating quantities based on measurements are intended for vector data handling. At the same time, the amount of complex-structured data, for example, XML structures or oriented acyclic graphs of knowledge bases in KIF, CycL languages, including representations of episodes in models of analogical reasoning [1–11] increase. However, methods of handling complex-structured data are computationally difficult. To increase handling efficiency, they are transformed into vectors, which allow us to estimate the characteristics important in the context of the problem being solved (for example, similarity measures) and to use these estimates to solve the original problem (see [7, 12] and references therein).

The complexity of handling vector data is proportional to the dimension and number of vectors. To deal with large arrays of large-dimensional vectors, a number of approaches are used. The most obvious is application of the brute force method, which is a powerful computing resource. For example, efficient implementation of vector operations is supported by vector (co)processors and, moreover, these operations are naturally parallelized. This allows using parallel computing tools: multi-core processors, parallel computers, and high-performance computing clusters and distributed computing systems (see [13–15] and references therein).

Another approach to increasing the efficiency of vector data handling is based on the fact that many applications do not require exact results of operations over original vectors (for example, calculation of scalar products, distances, etc.), an approximate but fast estimate is enough. For example, in similarity search problems it is often more important to quickly determine approximate "nearest neighbors" than to search for exact ones for a long time. To implement this approach, a number of methods of transformation of vector arrays is developed.

The methods reducing vector dimension [16, 17] can be adjusted for the special features of their basis used as a training sample. Methods of supervised and unsupervised training (adaptation) are applied.

Methods of unsupervised training detect the most informative (according to the given criteria) new coordinates in the data and reduce the dimension by eliminating some of them. The similarity measures of the original vectors can be distorted. For example, principal component analysis (PCA) reveals orthogonal directions in the data, which minimize the distortion of

International Scientific and Training Center of Information Technologies and Systems, National Academy of Sciences and Ministry of Education and Science of Ukraine, Kyiv, Ukraine, *dar@infrm.kiev.ua*. Translated from Kibernetika i Sistemnyi Analiz, No. 5, September–October, 2015, pp. 157–168. Original article submitted September 17, 2014.

the Euclidean distances between vectors. This reduces distance estimate, and the error increases with the number of eliminated measurements.

In a number of supervised dimension reduction methods, the similarity of resultant vectors reflects the “semantic similarity” of input data [18]. Obtaining this information involves a lot of labor on estimating the similarity of original objects, usually done by an expert.

A common disadvantage of many supervised dimension reduction methods is the computing complexity of the solution of their optimization problems, for example, singular value decomposition or gradient descent procedures (with possible local minima). Moreover, application of some methods involves difficulties in transformation of new data that were not used in training. Obtaining their reduced short representations can demand repeated training or estimation of the similarity with a considerable part of the training sample.

The shortcomings of supervised dimension reduction methods have caused the development of the approach to transformation of vector data without adaptation, so-called random projection [19–27]. In this method, to transform input vectors into output ones, multiplication by a random matrix is carried out; elements of the matrix are randomly generated and then fixed numbers from some distribution. In a number of studies (for example, [19, 20, 26]), the accuracy of estimates of some similarity–distinction measures is analyzed when random matrices of certain form are used. For example, it is shown in [26] that transformation by a simple (from the point of view of generation and application) binary random matrix with elements $\{0, 1\}$ allows estimating the Euclidean distance, norm, and scalar product of input vectors using output vectors.

Some approaches and methods require specialized algorithms for storage and handling of vector data in a certain format (representation). Sparse [28, 29] (with a small share of nonzero components) binary vectors are used, for example, in associative-projective neural networks [30, 31] and in the efficient binary version [32–34] of distributed associative memory [10, 11, 35–38].

The formation of binary output vectors with controlled sparsity, which can be used to estimate the similarity (angle value) of real-valued input vectors, is considered in [25] for a ternary random matrix with elements from $\{-1, 0, +1\}$, and in [27] for a binary matrix. In the present paper, we will analyze the estimates of the angle, Euclidean distance, and scalar product of input vectors based on a series of similarity characteristics of binary output vectors when a random binary matrix is used for the transformation.

## PROJECTION BY A RANDOM BINARY MATRIX AND ESTIMATE OF SIMILARITY MEASURES USING BINARY VECTORS

Similarly to [26, 27], we consider projection of vectors by a random binary matrix $\mathbf{R}$ with elements $r_{ij}$ from the set $\{0, 1\}$. Random variables (r.v.) whose realization are elements of $\mathbf{R}$ (ones and zeros) are independent and have identical distribution (i.i.d.). Each $r_{ij}$ takes the value 1 with probability $q$ and value 0 with probability $1-q$. Denote by $\mathbf{x}$ and $\mathbf{y}$ input real-valued vectors of dimension $D$ and by $\mathbf{u}=\mathbf{Rx}$ and $\mathbf{v}=\mathbf{Ry}$ the results of their projection (intermediate vectors of dimension $d$). Hence, the dimension of $\mathbf{R}$ is $(d \times D)$.

In the projection, each component $u_i$, $i = 1, \ldots, d$, of vector $\mathbf{u}$ is originally formed as the scalar product of the row $\mathbf{r}_i$ of matrix $\mathbf{R}$ by $\mathbf{x}$:

$$u_i = \langle \mathbf{r}_i, \mathbf{x} \rangle = \sum_{j=1}^{D} r_{ij} x_j. \tag{1}$$

The components $u_i$ are i.i.d. r.v. The expectation of $u_i$:

$$E\{u_i\} = E\left\{ \sum_{j=1}^{D} r_{ij} x_j \right\} = q \sum_{j=1}^{D} x_j \tag{2}$$

since $E\{r_{ij}\} = 1q + 0(1-q) = q$. The variance $u_i$:

$$V\{u_i\} = V\left\{ \sum_{j=1}^{D} r_{ij} x_j \right\} = \sum_{j=1}^{D} V\{r_{ij} x_j\} = \sum_{j=1}^{D} x_j^2 V\{r_{ij}\} = ||x||_2^2 \, (q - q^2) \tag{3}$$

since $r_{ij}$ is i.i.d., $E\{r_{ij}^2\} = 1^2 q + 0^2 (1-q) = q$, and $V(r_{ij}) = E\{r_{ij}^2\} - (E\{r_{ij}\})^2 = q - q^2$.

The standardization of $u_i$ is carried out by subtraction of expectation (2) $q\sum_{j=1}^{D} x_j$ (centering) and dividing by the root mean square deviation $||x||_2 \sqrt{q-q^2}$ (square root of (3)). The distribution of the standardized r.v. $u_i$ converges to the Gaussian one with zero mean value and unit variance. The convergence rate of $u_i$ is analyzed in [27] (for the review of the convergence problem for the general case see [39, 40]).

The binary output vector is formed by the binarizing threshold transformation of the intermediate vector $\mathbf{u} \to \mathbf{z}$:

$$z_i = 1 \text{ for } u_i > t_i; \; z_i = 0 \text{ otherwise, } i = 1, \dots, d, \tag{4}$$

where $t_i$ is the threshold value for the *i*th component of the output vector. We will use identical threshold values for all the components, $t_i \equiv t$. The given probability $p$ of the unit component $z$ of the binary output vector $\mathbf{z}$ is determined by choosing the appropriate threshold $t_p$. For a standardized r.v. $u$ with the Gaussian distribution

$$p(z=1) = p(u > t_p) = \frac{1}{\sqrt{2\pi}} \int_{t_p}^{\infty} e^{-\eta^2/2} d\eta = 1 - \Phi(t_p),$$

where $\Phi$ is the Gaussian cumulative distribution function. The value of $t_p$ to provide the necessary $p$ is defined as the quantile of the Gaussian distribution, corresponding to $1-p$. The vectors for $p < 0.5$ are sparse.

**Estimate of the Angle Between Input Vectors.** In [25], for the estimate of the angle $\theta$ between input vectors $\mathbf{x}$ and $\mathbf{y}$ it is proposed to use the estimates of the probability $p_{join}$ of the coincidence of unit components $z_{1,i} = 1$ and $z_{2,i} = 1$ in the vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ after the binarization $\mathbf{u} \to \mathbf{z}_1$ with threshold $t_1$ and $\mathbf{v} \to \mathbf{z}_2$ with threshold $t_2$. According to the multidimentional central limit theorem, joint distribution of r.v. $(u_i, v_i)$ converges to the two-dimensional Gaussian distribution. For standardized $(u_i, v_i)$, this probability can be found as a result of evaluation of the integral of the two-dimensional Gaussian distribution

$$p_{join} \equiv p(z_{1,i} = 1, \; z_{2,i} = 1 | \theta, t_1, t_2) = p(u_i > t_1, \; v_i > t_2 | \theta)$$

$$= \frac{1}{2\pi(1-\cos^2\theta)} \int_{t_1}^{\infty}\int_{t_2}^{\infty} \exp\left( -\frac{\eta_1^2 - 2\eta_1\eta_2\cos\theta + \eta_2^2}{2(1-\cos^2\theta)} \right) d\eta_1 d\eta_2. \tag{5}$$

Thus, the angle $\theta$ is related by the functional dependence with $p_{join}$: $\theta = f(p_{join})$, where $f$ is the function inverse to (5). Therefore, $\theta$ can be estimated as follows: tabulate (5); transform input vectors into output ones $\mathbf{z}_1$ and $\mathbf{z}_2$ using (1) and (4); estimate $p_{join}$ as $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$; in the table, find the value of $p_{join}$ nearest to $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$ and use the angle $\theta^*$ corresponding to it as the estimate of $\theta$.

Let us find the expectation and variance of the angle estimate by appliying the linearization of function of random argument [41]. Let there be an r.v. $\xi$ with known expectation $E\{\xi\} = E_\xi$ and variance $V\{\xi\} = V_\xi$, and let there be another r.v. $\zeta$, related to $\xi$ as $\zeta = \varphi(\xi)$, and in the neighbourhood of $E\{\xi\} = E_\xi$ function $\varphi$ be close to a linear one. Therefore, we can represent the relation between $\xi$ and $\zeta$ for small deviations from the mean value using a one-step Taylor series expansion as $\zeta \approx \varphi(E_\xi) + \varphi'(E_\xi)(\xi - E_\xi)$. Then we approximate the expectation and variance $\zeta$ as

$$E\{\zeta\} \approx \varphi(E_\xi) + \varphi'(E_\xi) E\{(\xi - E_\xi)\} = \varphi(E_\xi), \tag{6}$$

$$V\{\zeta\} \approx V\{\varphi(E_\xi)\} + (\varphi'(E_\xi))^2 V\{(\xi - E_\xi)\} = (\varphi'(E_\xi))^2 V_\xi. \tag{7}$$

The number $k = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle$ of coinciding unit components for the probability of their coincidence $p_{join}$ for different realizations of binary vectors of dimension $d$, as well as $p_{join}^* \equiv \langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$ and the angle estimate $\theta_{join}^* = f(p_{join}^*)$ are r.v. Applying (6) and (7) to $\theta_{join}^* = f(p_{join}^*)$, we obtain

$$E\{\theta_{join}^*\} = f(E\{p_{join}^*\}), \; V\{\theta_{join}^*\} = (f'(E\{p_{join}^*\}))^2 V\{p_{join}^*\}. \tag{8}$$

The random variable $k = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle$ has binomial distribution [25]: $P(k) = \beta(k, d, p_{join})$ with the expectation $E\{k\} = dp_{join}$ and variance $V\{k\} = dp_{join}(1 - p_{join})$. Therefore,

$$E\{p^*_{join}\} = E\{k/d\} = p_{join},$$

$$V\{p^*_{join}\} = V\{k/d\} = V\{k\}/d^2 = p_{join}(1 - p_{join})/d. \quad (9)$$

Taking into account (9), we write (8) as

$$E\{\theta^*_{join}\} = f_{join}(p_{join}),\ V\{\theta^*_{join}\} = (f'_{join}(p_{join}))^2 p_{join}(1 - p_{join})/d. \quad (10)$$

To estimate the value of the derivative $f'_{join}(p_{join}) \equiv \theta'(p_{join})$ using the known value of probability $p_{join}$, we can use the tabulated function (5), having found the nearest value in the table from $p_{join}$ and the angle corresponding to it $\theta$, and having calculated $\Delta\theta / \Delta p$, where $\Delta\theta$ and $\Delta p$ is the difference of the values of angles and probabilities between the value in the found and adjacent cells of the table, respectively. Hence, if $\theta$ is known, to find the "analytical" expectation and the variance of its estimate using $p^*_{join}$, we will find in the tabulated (5) using $\theta$ the respective probability $p_{join}$ and $f'(p_{join}) \equiv \theta'(p_{join})$ and use (10).

Along with the estimate of the angle with respect to the empirical probability $p_{join}$, we will analyze its estimate with respect to the empirical conditional probability $p_{cond}$ of the coincidence of unit components of output binary vectors, and also with respect to the empirical probability $p_{equ}$ of the coincidence of values of their components (both unit and zero).

Let us express $p_{cond}$ in terms of $p_{join}$. For binary vectors with identical probability $p$ of the unit component we have

$$p_{cond} \equiv p(z_{1,i} = 1 | z_{2,i} = 1) = p(z_{1,i} = 1,\ z_{2,i} = 1)/\ p(z_{2,i} = 1) \equiv p_{join}\ /\ p. \quad (11)$$

We will find the dependence of $p_{cond}$ on the angle by dividing (5) by $p$.

Having found $p^*_{cond} \equiv \langle \mathbf{z}_1, \mathbf{z}_2 \rangle / |\mathbf{z}_2|$, where $|\mathbf{z}_2|$ is the number of unit components of $\mathbf{z}_2$, we can calculate

$$E\{p^*_{cond}\} = E\{k/(pd)\} = p_{cond},$$

$$V\{p^*_{cond}\} = V\{k/(pd)\} = V\{k\}/(pd^2) = p_{cond}(1 - p_{cond})/(pd), \quad (12)$$

since no more than $pd$ unit components can coincide. We will find the values of $f'(p_{cond}) \equiv \theta'(p_{cond})$ from the tabulated $p_{cond}$ similarly to $p_{join}$ and obtain

$$E\{\theta^*_{cond}\} = f_{cond}(p_{cond}),$$

$$V\{\theta^*_{cond}\} = (f'_{cond}(p_{cond}))^2 p_{cond}(1 - p_{cond})/(pd). \quad (13)$$

For $p_{equ}$ we write

$$p_{equ} \equiv p(z_{1,i} = z_{2,i}) = 1 - p(z_{1,i} \neq z_{2,i}) = 1 - (p(z_{1,i} = 1)$$

$$+\ p(z_{2,i} = 1) - 2p(z_{2,i} = 1, z_{2,i} = 1)) = 1 - (2p - 2p_{join}). \quad (14)$$

We will find the dependence of $p_{equ}$ on the angle using $p_{join}$ (5) and $p$.

Having found $p^*_{equ} \equiv 1 - \mathbf{z}_1 \oplus \mathbf{z}_2\ /\ d$, where $\oplus$ is the component-wise "exclusive OR" operation, similarly to (9), (10), (12), and (13) we obtain

$$E\{p^*_{equ}\} = p_{cond},\ V\{p^*_{equ}\} = p_{equ}(1 - p_{equ})/d, \quad (15)$$

$$E\{\theta^*_{equ}\} = f_{equ}(p_{equ}),\ V\{\theta^*_{equ}\} = (f'_{equ}(p_{equ}))^2 p_{equ}(1 - p_{equ})/d. \quad (16)$$

As shown in [25], the probability $p_{join}$ of the coincidence of unit components of the output binary vectors monotonically decreases as the angle $\theta$ increases. According to (11) and (14), this is also true for $p_{cond}$ and $p_{equ}$. Hence, the estimates of these probabilities $p^*_{join}$, $p^*_{cond}$, and $p^*_{equ}$ can be useful as the measures of similarity of input vectors (without using them to calculate angle estimates).

**Estimate of the Scalar Product and Euclidean Distance.** As was mentioned above, the standardization of $u_i$ to obtain binary output vectors requires the Euclidean norms of original vectors $||\mathbf{x}||$ and $||\mathbf{y}||$ to be known. The norm is calculated once for one vector and the obtained number is saved. This information, together with the angle estimate $\theta^*$ with respect to binary output vectors considered above can be used to estimate the scalar product $\langle \mathbf{x},\mathbf{y}\rangle^*$ and the (squared) Euclidean distance $||\mathbf{x}-\mathbf{y}||^{2*}$:

$$\langle \mathbf{x},\mathbf{y}\rangle^* = ||\mathbf{x}||\,||\mathbf{y}||\cos\theta^*, \tag{17}$$

$$||\mathbf{x}-\mathbf{y}||^{2*} = ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2\,||\mathbf{x}||\,||\mathbf{y}||\cos\theta^*. \tag{18}$$

For the scalar product with the use of linearization (7) we have

$$V\{\langle \mathbf{x},\mathbf{y}\rangle^*\} = ||\mathbf{x}||^2\;||\mathbf{y}||^2\,V\{\cos\theta^*\}$$

$$= ||\mathbf{x}||^2\;||\mathbf{y}||^2\,V\{\theta^*\}\left(\frac{d\cos\theta}{d\theta}\right)^2 = ||\mathbf{x}||^2\;||\mathbf{y}||^2\,V\{\theta^*\}\sin^2\theta. \tag{19}$$

Here, $V\{\theta^*\}$ depends on the method of angle estimation (10), (13), (15), (16).

For the squared Euclidean distance with the use of (7) we obtain

$$V\{||\mathbf{x}-\mathbf{y}||^{2*}\} = 4\,||\mathbf{x}||^2\;||\mathbf{y}||^2\,V\{\cos\theta^*\}$$

$$= 4\,||\mathbf{x}||^2\;||\mathbf{y}||^2\,V\{\theta^*\}\sin^2\theta = 4V\{\langle \mathbf{x},\mathbf{y}\rangle^*\}. \tag{20}$$

## EXPERIMENTAL ANALYSIS

We analyzed the behavior of the error of estimate of the angle between input vectors, scalar product, and the Euclidean distance between them. The estimates were found from the output binary vectors of different sparsity. As the error of angle estimate, we used the variance $V$; the error of the estimate of scalar product and distance was measured by the variation coefficient $V^{1/2}/E$. The results of the dependence of the error of angle estimate on the dimension of input and output vectors are presented in [27], where the dependence of the error on the angle between input vectors was analyzed.

To transform input real-valued vectors **x** and **y** into intermediate ones **u** and **v**, random matrices with binary $\{0, 1\}$ and ternary $\{-1, 0, +1\}$ elements were used; the probability of a nonzero element of matrices is $q = \{0.5, 0.1\}$. The threshold value $t$ for standardized values of components of intermediate vectors $u_i$ was selected to maintain the probability of unit component $p = \{0.5, 0.1\}$ in output binary vectors, i.e., $t_p = \{0.0, 1.282\}$.

Input vectors with $D = 1000$ and output binary vectors with $d = 1000$ were used. The similarity was varied by the concatenation of vectors **a**, **b**, and **c**: $\mathbf{x} = (\mathbf{a}\ \mathbf{b})$ and $\mathbf{y} = (\mathbf{c}\ \mathbf{b})$ of different dimensions. For example, if the dimensions of **a** and **c** are zero, we obtain identical vectors **b** of dimension $d$, and if the dimension of **b** is zero, we obtain different vectors **a** and **c** of dimension $d$. The components of **a**, **b**, and **c** were generated randomly from uniform distribution in $[-D, +D]$. Since the dimensions of input vectors ($D = 1000$) are big, their norms are close. A typical value of the angle between random vectors is approximately 90°, i.e., random vectors are almost orthogonal. For random vectors with uniformly distributed positive components, the angle is usually 40° to 45°. For such vectors, the orthogonality can be attained for nonintersecting parts of vectors $\mathbf{x} = (\mathbf{a}\ 0)$ and $\mathbf{y} = (0\ \mathbf{b})$.

For the estimates of joint probability $p^*_{join}$, the value of $\langle \mathbf{z}_1, \mathbf{z}_2\rangle / d$ can exceed the greatest possible value $p^{\max}_{join} = p$. To ensure the correspondence of the experimental results to formulas (9)–(11), "mirror" transformation $p - (p^*_{join} - p)$ was used for the case $p^*_{join} > p^{\max}_{join}$, and the obtained value of the angle was considered negative.

The results were averaged over 10,000 realizations of a random matrix.

The experiments have shown that values of errors for the analyzed parameters are close for binary and ternary random matrices. Therefore, we will present the results for binary matrix.

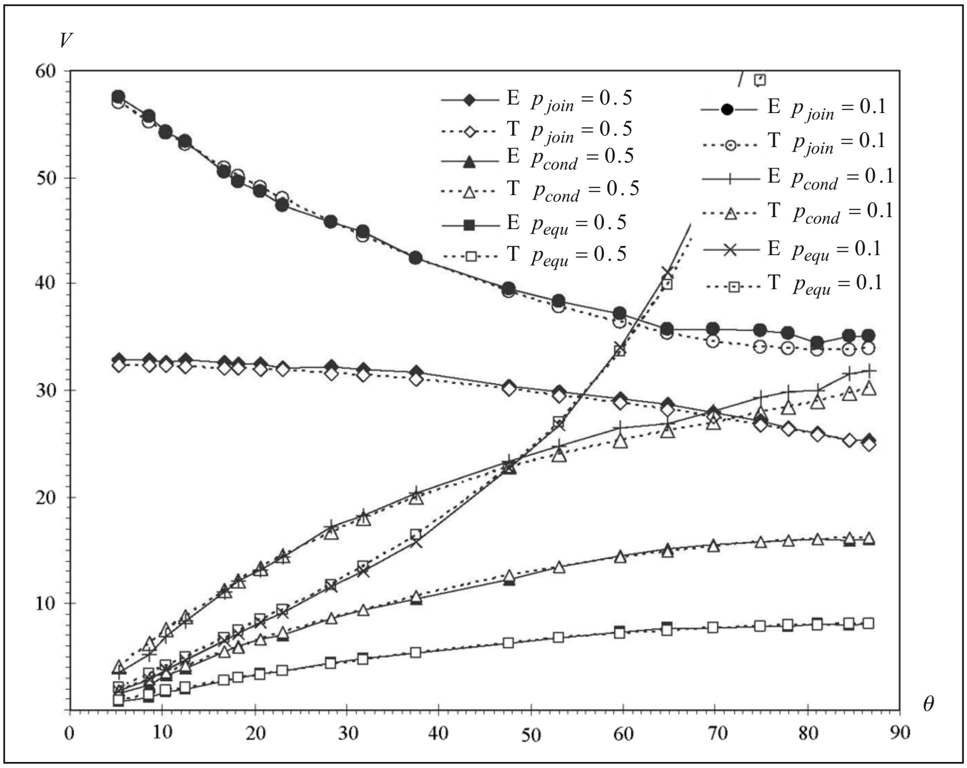

Fig. 1. Dependences of the error of estimate of the angle $\theta$ between output vectors on its value with $D = 1000$, $d = 1000$, and $q = 0.1$.

Figure 1 shows the dependences of error $V$ of the estimate of angle $\theta$ between the input vectors on its value. The estimates are found from $p^*_{join}$, $p^*_{cond}$, and $p^*_{equ}$; the notation T corresponds to the values obtained from (10), (13), and (16), and E to experimental results.

For $p = 0.5$, the analytical values of the errors of angle estimates are close to the respective experimental values. Angle estimates with respect to $p^*_{join}$ have the greatest error (its value decreases as the angle increases), with respect to $p^*_{equ}$ the least error, and with respect to $p^*_{cond}$ the error is intermediate. Zero error for zero angle is obtained for $p^*_{equ}$ and $p^*_{cond}$. The value of error grows with the angle. (Zero angle was not used in the above-mentioned experiments.)

As one would expect, for $p = 0.1$ the errors exceed the respective values for $p = 0.5$. The analytical values of the errors of angle estimates are close to the experimental ones corresponding to them for $p^*_{cond}$ and $p^*_{join}$ in the whole range of angles. The error of angle estimate with respect to $p^*_{cond}$ is still less than the error with respect to $p^*_{join}$ with the greatest difference for 0° and with approximately equal values for 90°. The values of the error of angle estimate with respect to $p^*_{equ}$ strongly differ from the case $p = 0.5$: if the angle is greater than 50°, its value exceeds the error with respect to $p^*_{cond}$, and if the angle is greater than 60°, it exceeds the error with respect to $p^*_{join}$; moreover, the analytical values of the error become less than experimental ones, i.e., approximation (7) does not hold. Thus, the angle estimate with respect to $p^*_{cond}$ is preferable, and the range of applicability of estimates with respect to $p^*_{equ}$ is restricted by small angles and decreases as the share of unit components in output vectors decreases.

Figure 2 shows the dependences of the error $V^{1/2} / E$ (variation coefficient) of the estimate of the scalar product of input vectors with respect to (17) on the value of angle $\theta$ for the angle estimates obtained with respect to $p^*_{join}$, $p^*_{cond}$, and $p^*_{equ}$. The notation T corresponds to the values of error obtained from (19) and from the angle estimates from (10), (13), and (16), and E corresponds to the experimental results.

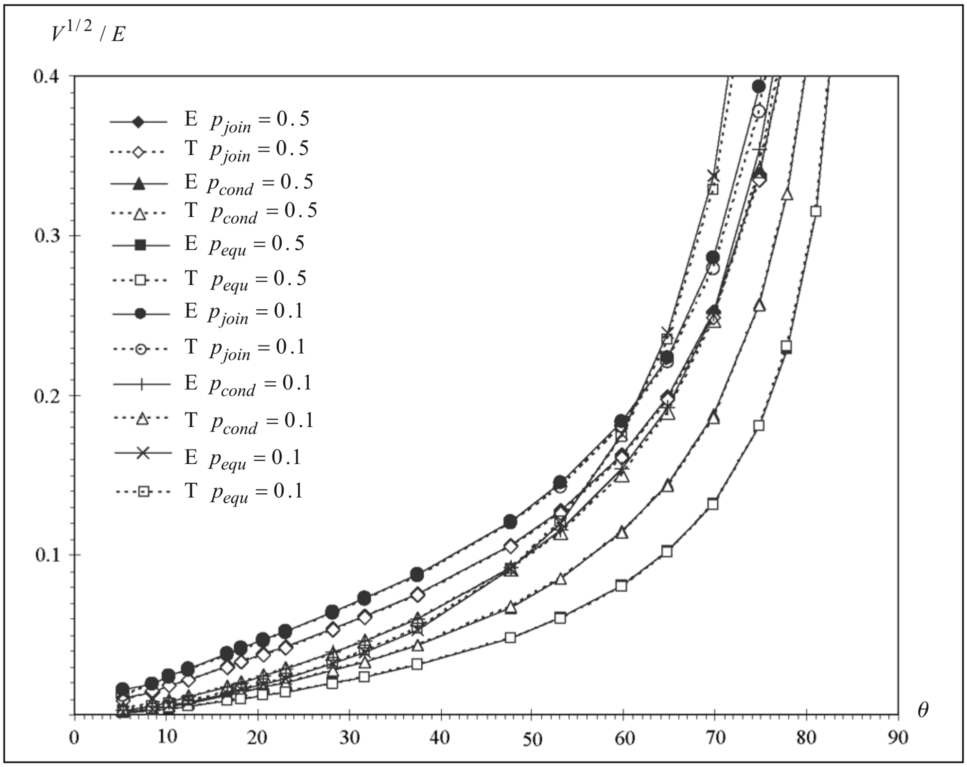

Fig. 2. Dependence of the error of estimate of scalar product on the angle $\theta$ between vectors with $D = 1000$, $d = 1000$, and $q = 0.1$.

For all the analyzed combinations of parameters, the values of errors of the estimate of scalar product for $p = 0.1$ exceeds the respective values for $p = 0.5$; as the angle between input vectors increases, the error grows especially fast if the angle is greater than 50°. Similarly to the angle estimate, the estimate with respect to $p^*_{equ}$ also has the least error for $p = 0.5$, but for $p = 0.1$ the error becomes the greatest if the angle is greater than 60°. For greater angle values, a difference between analytical and experimental values of errors for $p^*_{equ}$ is noticeable, for other cases these results are close. Thus, the greatest accuracy of the estimate of the scalar product is attained for small values of angles between input vectors: for $p = 0.5$ it is necessary to use the estimate with respect to $p^*_{equ}$, and for sparse output binary vectors the best results are attained for the estimate with respect to $p^*_{cond}$.

Figure 3 shows the dependences of the error $V^{1/2} / E$ (variation coefficient) of the estimate of squared Euclidean distance between input vectors with respect to (18) on the value of angle $\theta$, the analytical values of the error are obtained from (20); the notation in the description is similar to that in Figs. 1 and 2.

As well as for angle estimates and scalar product, the values of error estimates of the squared distance for $p = 0.1$ exceed the respective values for $p = 0.5$. However, the error does not grow but decreases as the angle increases. Since the distance between angles grows with the angle between vectors, the value of the relative error analyzed in the paper decreases, and this compensates for the growth of the variance of angle as the angle increases for $p^*_{cond}$ and $p^*_{equ}$. As a result, the error estimate of the squared distance with respect to $p^*_{cond}$ (and even with respect to $p^*_{equ}$) varies insignificantly in all the analyzed range of angles. The greatest error (and discrepancy between the analytical and experimental results) can be obtained for $p^*_{join}$ for small angles since the variance of angle estimate is maximum for this case. The variation coefficient is unstable in the neighborhood of zero, and hence it cannot be applied for an adequate error estimate for angles close to zero. As we mentioned above, the error estimate of angle with respect to $p^*_{equ}$ grows very fast as the angle approaches to 90° for sparse binary vectors.

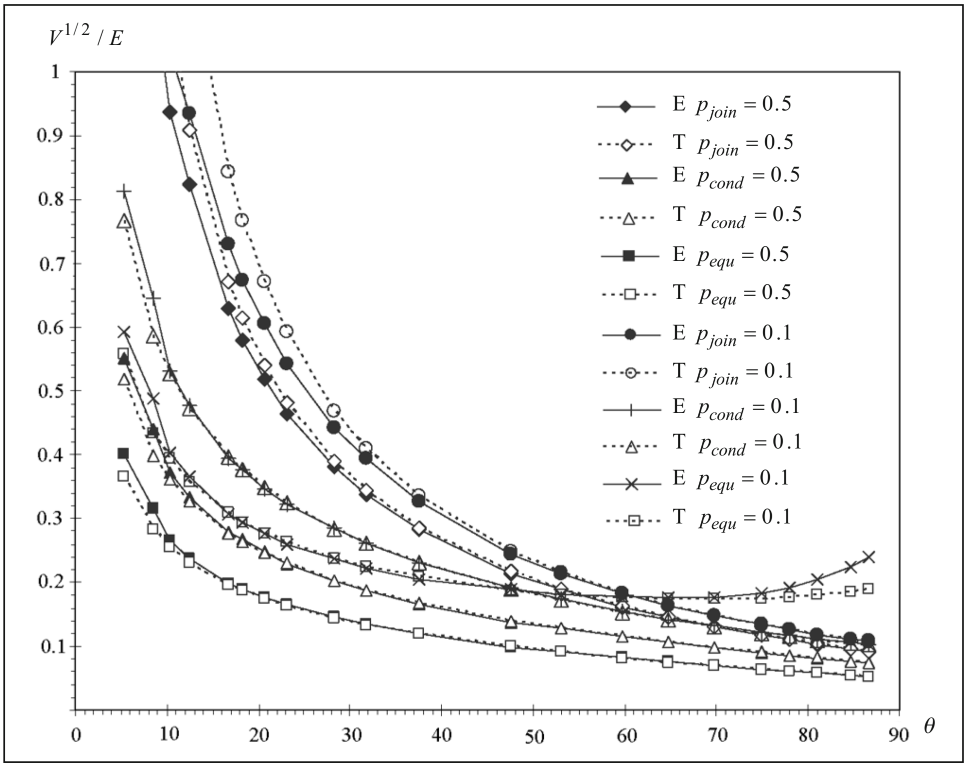

Fig. 3. Dependence of the error of estimate of the squared Euclidean distance on the angle $\theta$ between vectors with $D = 1000$, $d = 1000$, and $q = 0.1$.

Thus, for $p = 0.5$ estimates of the distance with respect to $p^*_{equ}$ have the least error; for sparse output binary vectors, the best results are attained for estimates with respect to $p^*_{cond}$ and with respect to $p^*_{equ}$ (for $p^*_{equ}$ if the angle is not too close to 90°); for small angle the distance error estimate grows fast.

For estimates of the angle between input vectors with respect to output binary vectors and for scalar product and Euclidean distance between input vectors with respect to the obtained angle estimate and known norms of input vectors in their application in similarity search problems, it is of interest to analyze a hybrid approach, where the similarity is estimated based on the angle for small angles and on the distance for large ones.

## CONCLUSIONS

We have analyzed the estimates of similarity measures of real-valued vectors with respect to binary ones, obtained by projection by a random binary matrix with elements $\{0, +1\}$ and output threshold transformation, which allows the control of the sparsity (share of unit components) of binary vectors. The similarity of the latter was estimated using the measures based on (normed to their dimension $d$) scalar product $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d \equiv |\mathbf{z}_1 \wedge \mathbf{z}_2| / d$ and related $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / |\mathbf{z}_2|$ and $1 - \mathbf{z}_1 \oplus \mathbf{z}_2 / d = 1 - (|\mathbf{z}_1| + |\mathbf{z}_2| - 2|\mathbf{z}_1 \wedge \mathbf{z}_2|)/d$, where $\oplus$ and $\wedge$ are the operations of component-wise XOR and AND, respectively, and $|\mathbf{z}|$ is the number of unit components in the binary vector $\mathbf{z}$. These measures are natural estimates of the similarity of the original real-valued vectors $\mathbf{x}$ and $\mathbf{y}$: their values monotonically decrease as the angle $\theta$ between the vectors increases and allow estimating its value. The estimate of angle $\theta$ for known values of the Euclidean norms $||\mathbf{x}||$ and $||\mathbf{y}||$ has also allowed estimating the scalar product $\langle \mathbf{x}, \mathbf{y} \rangle^* = ||\mathbf{x}|| \, ||\mathbf{y}|| \cos \theta^*$ and the Euclidean distance $||\mathbf{x} - \mathbf{y}||^* = (||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2 ||\mathbf{x}|| \, ||\mathbf{y}|| \cos \theta^* )^{1/2}$.

We have analyzed, analytically and experimentally, the dependences of error of the estimate of angle, scalar product, and Euclidean distance between input real-valued vectors on the angle between them. The dependences substantially differ for the analyzed three similarity measures of binary vectors and for estimates of the scalar product and Euclidean distance. This allows using the estimates with the least error in different ranges of angles.

For binary and ternary random matrices, the values of error are close for the analyzed values of parameters. However, implementing the transformation by means of a binary random matrix is computationally simpler.

A promising subject of the further studies is the analysis of the influence of modifications of the proposed methods on the accuracy of similarity estimates, for example: instead of the probability of unit element in the matrix, using the actual share of unit elements in the entire matrix and in its rows and columns; using random matrices with fixed number of randomly arranged unit elements in the entire matrix and in its rows and columns; taking into account the actual share of unit components in output binary vectors in the angle estimate based on their scalar product.

Output binary vectors, where not the semantics of separate components is important but their reflection of the similarity of represented objects, are an example of randomized distributed representations [1–11, 30, 31, 35, 36, 42–45]. Distributed representations can also be formed in inner layers of multilayer networks in the course of training [46–48]. Distributed representations are used to represent semantic similarity [18, 42, 49–51], sequences [29–31, 43, 51–54], complex hierarchically structured objects [1, 2, 4–11, 30, 31, 35, 36, 42–45, 55–57] required for models and systems of artificial intelligence [3, 31, 35, 36, 57–59].

## REFERENCES

1. D. A. Rachkovskij, “Representation and processing of structures with binary sparse distributed codes,” IEEE Trans. on Knowledge and Data Engineering, **13**, No. 2, 261–276 (2001).
2. D. A. Rachkovskij, “Some approaches to analogical mapping with structure sensitive distributed representations,” J. Experim. and Theor. Artificial Intelligence, **16**, No. 3, 125–145 (2004).
3. M. Stanojevic and S. Vranes, “Semantic approach to knowledge processing,” WSEAS Trans. on Inform. Sci. and Appl., **5**, No. 6, 913–922 (2008).
4. S. V. Slipchenko and D. A. Rachkovskij, “Analogical mapping using similarity of binary distributed representations,” Intern. J. Inform. Theories and Applications, **16**, No. 3, 269–290 (2009).
5. R. W. Gayler and S. D. Levy, “A distributed basis for analogical mapping,” in: Proc. 2nd Intern. Analogy Conference, NBU Press, Sofia (2009), pp. 165–174.
6. D. A. Rachkovskij and S. V. Slipchenko, “Similarity-based retrieval with structure-sensitive sparse binary distributed representations,” Computational Intelligence, **28**, No. 1, 106–129 (2012).
7. V. I. Gritsenko, D. A. Rachkovskij, A. D. Goltsev, V. V. Lukovych, I. S. Misuno, E. G. Revunova, S. V. Slipchenko, A. M. Sokolov, and S. A. Talayev, “Neural distributed representation for intelligent information technologies and modeling of thinking,” Cybernetics and Computer Engineering, **173**, 7–24 (2013).
8. M. Pickett and D. Aha, “Using cortically-inspired algorithms for analogical learning and reasoning,” Biologically Inspired Cognitive Architectures, **6**, 76–86 (2013).
9. B. Emruli, R. W. Gayler, and F. Sandin, “Analogical mapping and inference with binary spatter codes and sparse distributed memory,” Intern. Joint Conference on Neural Networks (IJCNN), 4–9 Aug. 2013, Dallas, TX, IEEE, (2013), pp. 1–8.
10. B. Emruli and F. Sandin, “Analogical mapping with sparse distributed memory: A simple model that learns to generalize from examples,” Cognitive Computation, **6**, No. 1, 74–88 (2014).
11. D. Widdows and T. Cohen, “Reasoning with vectors: A continuous model for fast robust inference,” Logic J. of the IGPL, **23**, No. 2, 141–173 (2015).
12. P. Indyk and J. Matousek, “Low-distortion embeddings of finite metric spaces,” in: J. E. Goodman and J. O’Rourke (eds.), Handbook of Discrete and Computational Geometry, Discrete Mathematics and its Applications, Chapman & Hall/CRC, Boca Raton (2004), pp. 177–196.
13. S. R. Upadhyaya, “Parallel approaches to machine learning — A comprehensive survey,” J. of Parallel and Distributed Computing, **73**, No. 3, 284–292 (2013).
14. N. Kussul, A. Shelestov, S. Skakun, and O. Kravchenko, “High-performance intelligent computations for environmental and disaster monitoring,” Intern. J. Inform. Technologies and Knowledge, **3**, No. 2, 135–156 (2009).
15. N. Kussul, A. Shelestov, and S. Skakun, “Grid technologies for satellite data processing and management within international disaster monitoring projects,” in: S. Fiore and G. Aloisio (eds.), Grid and Cloud Database Management, Springer-Verlag, Berlin–Heidelberg (2011), pp. 279–306.

16. L. J. P. Van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," in: Tech. Rep. TiCC-TR 2009-005, Tilburg Centre Creative Comput., Tilburg Univ., Tilburg, (2009).

17. C. J. C. Burges, "Dimension reduction: A guided tour," Foundations and Trends in Machine Learning, **2**, No. 4, 275–365 (2010).

18. A. Sokolov and S. Riezler, "Task-driven greedy learning of feature hashing functions," in: Proc. NIPS'13 Workshop, Big Learning: Advances in Algorithms and Data Management, Lake Tahoe, (2013), pp. 1–5.

19. S. S. Vempala, The Random Projection Method, American Mathematical Society, Providence (2004).

20. P. Li, T. J. Hastie, and K. W. Church, "Very sparse random projections," in: 12th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining, ACM Press, Philadelphia (2006), pp. 287–296.

21. E. G. Revunova and D. A. Rachkovskij, "Using randomized algorithms for solving discrete ill-posed problems," Inform. Theories and Applications, **16**, No. 2, 176–192 (2009).

22. E. G. Revunova, "Study of error components for solution of the inverse problem using random projections," Mathematical Machines and Systems, No. 4, 33–42 (2010).

23. E. G. Revunova and D. A. Rachkovskij, "Stable transformation of a linear system output to the output of system with a given basis by random projections," in: The 5th International Workshop on Inductive Modelling, IWIM (2012), pp. 37–41.

24. D. A. Rachkovskij and E. G. Revunova, "Randomized method for solving discrete ill-posed problems," Cybern. Syst. Analysis, **48**, No. 4, 621–635 (2012).

25. D. A. Rachkovskij, I. S. Misuno, and S. V. Slipchenko, "Randomized projective methods for construction of binary sparse vector representations," Cybern. Syst. Analysis, **48**, No. 1, 146–156 (2012).

26. D. A. Rachkovskij, "Vector data transformation with random binary matrices," Cybern. Syst. Analysis, **50**, No. 6, 960–968 (2014).

27. D. A. Rachkovskij, "Formation of similarity-reflecting binary vectors with random binary projections," Cybern. Syst. Analysis, **51**, No. 2, 313–323 (2015).

28. G. Rinkus, "Quantum computation via sparse distributed representation," NeuroQuantology, **10**, No. 2, 311–315 (2012).

29. G. J. Rinkus, "Sparsey™: Event recognition via deep hierarchical sparse distributed codes," Frontiers in Computational Neuroscience, **8**, Article 160, 1–44 (2014).

30. E. M. Kussul and D. A. Rachkovskij, "Multilevel assembly neural architecture and processing of sequences," in: A. V. Holden and V. I. Kryukov (eds.), Neurocomputers and Attention: Vol. II, Connectionism and Neurocomputers, Manchester University Press, Manchester–New York (1991), pp. 577–590.

31. D. A. Rachkovskij, E.M. Kussul, and T. N. Baidyk, "Building a world model with structure-sensitive sparse binary distributed representations," Biologically Inspired Cognitive Architectures, **3**, 64–86 (2013).

32. A. Kartashov, A. Frolov, A. Goltsev, and R. Folk, "Quality and efficiency of retrieval for Willshaw-like autoassociative networks: III. Willshaw-Potts model," Network: Computation in Neural Systems, **8**, No. 1, 71–86 (1997).

33. A. A. Frolov, D. A. Rachkovskij, and D. Husek, "On information characteristics of Willshaw-like auto-associative memory," Neural Network World, **12**, No. 2, 141–158 (2002).

34. A. A. Frolov, D. Husek, and D. A. Rachkovskij, "Time of searching for similar binary vectors in associative memory," Cybern. Syst. Analysis, **42**, No. 5, 615–623 (2006).

35. D. Kleyko, E. Osipov, A. Senior, A. I. Khan, and Y. A. Sekercioglu, "Holographic graph neuron: A bio-inspired architecture for pattern processing" (2015), http://arxiv.org/pdf/1501.03784v1.pdf.

36. B. Emruli, F. Sandin, and J. Delsing, "Vector space architecture for emergent interoperability of systems by learning from demonstration," Biologically Inspired Cognitive Architectures, **11**, 53–64 (2015).

37. D. W. Nowicki and O. K. Dekhtyarenko, "Averaging on Riemannian manifolds and unsupervised learning using neural associative memory," in: Proc. ESANN 2005, Bruges, Belgium, April, 27–29 (2005), pp. 181–189.

38. A. Knoblauch, G. Palm, and F. T. Sommer, "Memory capacities for synaptic and structural plasticity," Neural Computation, **22**, No. 2, 289–341 (2010).

39. V. Korolev and I. Shevtsova, "An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums," Scandinavian Actuarial J., No. 2, 81–105 (2012).

40. I. G. Shevtsova, "On the absolute constants in the Berry–Esseen-type inequalities," in: Doklady Mathem., **89**, No. 3, 378–381 (2014).

41. E. S. Ventsel', Probability Theory [in Russian], Nauka, Moscow (1969).

42. D. Widdows and T. Cohen, "Real, complex, and binary semantic vectors," Lecture Notes in Computer Science, **7620** (2012), pp. 24–35.

43. R. S. Omelchenko, "Spellchecking software on the basis of distributed representations," Problemy Programmir., No. 4, 35–42 (2013).

44. T. Cohen, D. Widdows, M. Wahle, and R. Schvaneveldt, "Orthogonality and orthography: Introducing measured distance into semantic space," Lecture Notes in Computer Science, **8369** (2014), pp. 34–46.

45. P. Kanerva, G. Sjodin, J. Kristoferson, R. Karlsson, B. Levin, A. Holst , J. Karlgren, and M. Sahlgren, "Computing with large random patterns," in: Foundations of Real-World Intelligence, CSLI Publications, Stanford (2001), pp. 251–311.

46. A. M. Reznik, A. A. Galinskaya, O. K. Dekhtyarenko, and D. W. Nowicki, "Preprocessing of matrix QCM sensors data for the classification by means of neural network," Sensors and Actuators B, **106**, 158–163 (2005).

47. A. N. Chernodub, "Direct method for training feed-forward neural networks using batch extended Kalman filter for multi-step-ahead predictions," Lecture Notes in Computer Science, **8131** (2013), pp. 138–145.

48. A. N. Chernodub, "Training neural networks for classification using the extended Kalman filter: A comparative study," Optical Memory and Neural Networks, **23**, No. 2, 96–103 (2014).

49. I. S. Misuno, D. A. Rachkovskij, and S. V. Slipchenko, "Vector and distributed representations reflecting semantic relatedness of words," Mathematical Machines and Systems, No. 3, 50–67 (2005).

50. A. Sokolov, "LIMSI: Learning semantic similarity by selecting random word subsets," in: Proc. 6th Intern. Workshop on Semantic Evaluation (SEMEVAL'12), Association for Computational Linguistics (2012), pp. 543–546.

51. A. Sokolov, "Vector representations for efficient comparison and search for similar strings," Cybern. Syst. Analysis, **43**, No. 4, pp. 484–498 (2007).

52. D. Kleyko and E. Osipov, "On bidirectional transitions between localist and distributed representations: The case of common substrings search using vector symbolic architecture," Procedia Computer Science, **41**, 104–113 (2014).

53. O. Rasanen and S. Kakouros, "Modeling dependencies in multiple parallel data streams with hyperdimensional computing," Signal Processing Letters, IEEE, **21**, No. 7, 899–903 (2014).

54. G. L. Recchia, M. Sahlgren, P. Kanerva, and M. N. Jones, "Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation," Computational Intelligence and Neuroscience, Article ID 986574 (2015).

55. V. Kvasnicka and J. Pospichal, "Deductive rules in holographic reduced representation," Neurocomputing, **69**, 2127–2139 (2006).

56. S. I. Gallant and T. W. Okaywe, "Representing objects, relations, and sequences," Neural Computation, **25**, No. 8, pp. 2038–2078 (2013).

57. F. Sandin, A. I. Khan, A. G. Dyer, A. H. M. Amin, G. Indiveri, E. Chicca, and E. Osipov, "Concept learning in neuromorphic vision systems: What can we learn from insects?" J. Software Eng. and Applic., **7**, No. 5 387–395 (2014).

58. A. A. Letichevsky, "Theory of interaction, insertion modeling, and cognitive architectures," Biologically Inspired Cognitive Architectures, **8**, 19–32 (2014).

59. A. A. Letichevsky, O. O. Letychevskyi, V. S. Peschanenko, and A. A. Huba, "Generating symbolic traces in the insertion modeling system," Cybern. Syst. Analysis, **51**, No. 1, 5–15 (2015).