Data-reducing principal component analysis (PCA) is NP-hard even under the simplest interval uncertainty

M. Koshelev
Baylor College of Medicine, Houston, USA
e-mail: misha680hnl@gmail.com

Principal component analysis (PCA) is one of the most widely used methods for reducing the data size. In practice, data is known with uncertainty, so we need to apply PCA to this uncertain data. Several authors developed algorithms for PCA under interval uncertainty. It is known that in general, the problem of PCA under interval uncertainty is NP-hard.

The usual NP-hardness proof uses situations in which all measurement results come with interval uncertainty. In practice, often, most measurements are reasonably accurate, and only a few (or even one) variables are measured with significant uncertainty. When we consider such situations, will the PCA still be NP-hard?

In this paper, we prove that even in the simplest case when for each object, at most one data points comes with interval uncertainty, the PCA problem is still NP-hard.

Keywords: principal component analysis, interval uncertainty, NP-hard.

1. Data reduction, PCA, and interval uncertainty: a brief reminder

Need to reduce the data size. In many real-life situations, for each object and/or situation k, we measure a large number d of variables. As a result of these measurements, we get the values $x_{k,1}, \ldots, x_{k,d}$ corresponding to different objects $k = 1, \ldots, n$. When the number of variables d is large, processing all this data requires a lot of computation time.

Example. Such a large amount of data occurs in 3-D medical imaging. For example, in functional magnetic resonance imaging (fMRI), for each of many patients, we measure the intensity values at dozens of thousands of voxels at dozens of moments of time; see, e. g. [3, 4, 15, 17]. As a result, processing all this data requires a large amount of time-consuming computations.

Possibility to reduce the data size. Often, the measured values are strongly dependent on each other. In such situations, it is possible to use this dependence to reduce the data size.

Principal component analysis (PCA): a brief reminder. For the case of linear dependence, the technique for correspondingly reducing the size of the data is called *principal component analysis* (PCA, for short; see, e. g., [18]). This technique was first invented by the famous statistician K. Pearson in the early 20 century [28].

The use of the original data means, in effect, that we represent each data vector $x_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,i}, \ldots, x_{k,d})$ as a linear combination of the basis vectors $u_1 = (1, 0, \ldots, 0)$, $u_2 = (0, 1, 0, \ldots, 0)$, ..., $u_i = (0, \ldots, 0, 1, 0, \ldots, 0)$, ..., $u_d = (0, \ldots, 0, 1)$:

$$x_k = x_{k,1} \cdot u_1 + x_{k,2} \cdot u_2 + \ldots + x_{k,i} \cdot u_i + \ldots + x_{k,d} \cdot u_d. \quad (1)$$

The basis vectors u_i are *orthonomal* in the sense that different vectors are orthogonal, i. e., $\langle u_i, u_j\rangle = 0$ for $i \neq j$, where

$$\langle a, b\rangle \overset{\text{def}}{=} a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_i \cdot b_i + \ldots + a_d \cdot b_d, \tag{2}$$

and each of these vectors has a unit Euclidean norm $\|u_i\|_2 = 1$, where for every vector $a = (a_1, \ldots, a_d)$, its Euclidean norm $\|a\|_2$ is defined by the formula

$$\|a\|_2^2 \overset{\text{def}}{=} \langle a, a\rangle = a_1^2 + a_2^2 + \ldots + a_i^2 + \ldots + a_d^2. \tag{3}$$

The main idea behind PCA is that instead of using the standard orthonormal basis, we find a different orthonormal basis $e_i = (e_{i,1}, \ldots, e_{i,d})$ for which $\langle e_i, e_j\rangle = 0$ for $i \neq j$ and $e_i^2 = \langle e_i, e_i\rangle = 1$. With respect to this basis, each data vector x_k can be represented as

$$x_k = y_{k,1} \cdot e_1 + y_{k,2} \cdot e_2 + \ldots + y_{k,i} \cdot e_i + \ldots + y_{k,d} \cdot e_d, \tag{4}$$

where, due to orthonormality, we have, for every i,

$$\langle e_i, x_k\rangle = y_{k,1} \cdot \langle e_i, e_1\rangle + y_{k,2} \cdot \langle e_i, e_2\rangle + \ldots + y_{k,i-1} \cdot \langle e_i, e_{i-1}\rangle + y_{k,i} \cdot \langle e_i, e_i\rangle +$$

$$+ y_{k,i+1} \cdot \langle e_i, e_{i+1}\rangle + \ldots + y_{k,d} \cdot \langle e_i, e_d\rangle =$$

$$= y_{k,1} \cdot 0 + y_{k,2} \cdot 0 + \ldots + y_{k,i-1} \cdot 0 + y_{k,i} \cdot 1 + y_{k,i+1} \cdot 0 + \ldots + y_{k,d} \cdot 0 = y_{k,i}, \tag{5}$$

hence

$$y_{k,i} = \langle e_i, x_k\rangle = e_{i,1} \cdot x_{k,1} + e_{i,2} \cdot x_{k,2} + \ldots + e_{i,j} \cdot x_{k,j} + \ldots + e_{i,d} \cdot x_{k,d}. \tag{6}$$

Then, for each data point x_k, we only use the first $p < d$ values $y_{k,1}, \ldots, y_{k,p}$.

As a result, instead of the original vector (4), we use an approximate value

$$X_k = y_{k,1} \cdot e_1 + y_{k,2} \cdot e_2 + \ldots + y_{k,i} \cdot e_i + \ldots + y_{k,p} \cdot e_p. \tag{7}$$

We want to select the vectors $e_1, \ldots, e_p$ for which $X_k \approx x_k$ for all objects $k = 1, \ldots, n$, i. e., for which $X_{ki} \approx x_{ki}$ for all objects k and for all variables i.

The values x_{ki} form a $(n \cdot d)$-dimensional vector x. Similarly, the values X_{ki} form a $(n \cdot d)$-dimensional vector X. We want each coordinate $x_{k,i}$ of the vector x to be close to the corresponding coordinate of the vector X. In other words, we want the approximation vector X to be as close to the original data vector x as possible. A reasonable measure of distance between the two vectors is the Euclidean distance

$$\|X - x\|_2 \overset{\text{def}}{=} \sqrt{\sum_{k=1}^{n} \sum_{i=1}^{d} (X_{k,i} - x_{k,i})^2}. \tag{8}$$

Thus, we should select the vectors $e_1, \ldots, e_p$ for which this distance $\|X - x\|_2$ is the smallest possible.

This minimization formulation can be simplified if we take into account that the square root is a strictly increasing function and thus, minimizing the square root is equivalent to minimizing the sum of the squares

$$\|X - x\|_2^2 = \sum_{k=1}^{n} \sum_{i=1}^{d} (X_{k,i} - x_{k,i})^2. \tag{9}$$

Here, by the definition of the Euclidean norm,

$$\sum_{i=1}^{d}(X_{k,i} - x_{k,i})^2 = \|X_k - x_k\|_2^2, \tag{10}$$

so we arrive at the following precise formulation.

Select the vectors $e_1, \ldots, e_p$ in such a way that the mean squared difference between the original data vectors x_k and the approximate vectors $\widetilde{x}_k$ is the smallest possible:

$$\text{minimize} \ \ \|X_1 - x_1\|_2^2 + \|X_2 - x_2\|_2^2 + \ldots + \|X_k - x_k\|_2^2 + \ldots + \|X_n - x_n\|_2^2. \tag{11}$$

Already Pearson showed that this minimum is attained if we take, as $e_1, \ldots, e_p$, the eigenvectors of the covariance matrix

$$C_{i,j} \stackrel{\text{def}}{=} x_{1,i} \cdot x_{1,j} + x_{2,i} \cdot x_{2,j} + \ldots + x_{k,i} \cdot x_{k,j} + \ldots + x_{n,i} \cdot x_{n,j} \tag{12}$$

that correspond to the p largest eigenvalues.

Comment. It should be mentioned that the same PCA technique is also used when we have a reasonably small data size d. In such situations, PCA is used to solve a *different* practical problem: namely, to find appropriate *factors*, i. e., combinations of variables which are the most relevant for a given process.

In this paper, however, we are mainly in the data-reducing applications of PCA.

Need to take interval uncertainty into account. In practice, measurements are never absolutely exact. In general, the measured values $\widetilde{x}_{k,i}$ are different from the actual (unknown) values $x_{k,i}$. In other words, the measurement inaccuracy is usually non-zero:

$$\Delta x_{k,i} \stackrel{\text{def}}{=} \widetilde{x}_{k,i} - x_{k,i} \neq 0. \tag{13}$$

In some cases, we know the probability distribution for the measurement inaccuracies $\Delta x_{k,i}$. However, frequently, we do not know this probability distribution. Often, the only information that we have about the measurement inaccuracy $\Delta x_{k,i}$ is the upper bound $\Delta_{k,i}$ on its absolute value:

$$|\Delta x_{k,i}| \le \Delta_{k,i}. \tag{14}$$

After each such measurement, the only information that we have about $x_{k,i}$ is that this belongs to the interval

$$x_{k,i} \in \mathbf{x}_{k,i} = [\widetilde{x}_{k,i} - \Delta_{k,i}, \widetilde{x}_{k,i} + \Delta_{k,i}]. \tag{15}$$

In other words, we get an interval uncertainty (see, e. g., [25]). We need to take interval uncertainty into account when we use PCA to reduce the data size.

2. Data-reducing PCA under interval uncertainty: what is known

PCA under interval uncertainty: known algorithms. The need for PCA under interval uncertainty is well known. There exist several efficient algorithms for PCA under interval uncertainty; see, e. g., [1, 2, 6–8, 11–14, 16, 21–26, 30, 31] and references therein.

Most of these algorithms aim at the factor applications of PCA, but they can be used in data reduction as well.

Data-reducing PCA under interval uncertainty: towards a precise formulation of the problem. In data reduction, our objective is to decrease the size of the data as much as possible. In the usual PCA, we select the basis vectors $e_1, \ldots, e_p$ for which the corresponding sum of the squares (11) is the smallest possible.

Because of the interval uncertainty, we can now also select the values $x_{k,i}$ within the corresponding intervals.

For example, suppose that for almost all objects k, we know the exact values of $x_{k,i}$, and these exact values satisfy the property $x_{k,2} = x_{k,1}$. This means that the second quantity is redundant — and we can therefore reduce the data size by keeping only the values $x_{k,1}$.

This redundancy may not survive when we get more data — but as long as the assumption $x_{k,2} = x_{k,1}$ is confirmed by all the known data, it makes sense to use this assumption to reduce the data.

Suppose now that we now add, to that data, a new object k_0 for which the values $x_{k_0,1}$ and $x_{k_0,2}$ are only known with interval uncertainty, i. e., for which, instead of the actual values $x_{k_0,i}$ we only know the intervals $\mathbf{x}_{k_0,1}$ and $\mathbf{x}_{k_0,2}$ of possible values. If these two intervals have a common point, this means that the new data is still consistent with the assumption that $x_{k,2} = x_{k,1}$ — and thus, it still makes sense to use this assumption to reduce the data.

This example shows that it is reasonable to select *both* the vectors e_i and the values $x_{k,i} \in \mathbf{x}_{k,i}$ for which the approximation is the best. Thus, we arrive at the following formulation.

Data-reducing PCA under interval uncertainty: a precise formulation of the problem. We are given intervals $\mathbf{x}_{k,i}$. We need to select the vectors $e_1, \ldots, e_p$ *and* the values $x_{k,i} \in \mathbf{x}_{k,i}$ in such a way that the mean square difference between the original data vectors x_k and the approximate vectors $\widetilde{x}_k$ is the smallest possible:

$$\text{minimize} \quad \|X_1 - x_1\|_2^2 + \|X_2 - x_2\|_2^2 + \ldots + \|X_k - x_k\|_2^2 + \ldots + \|X_n - x_n\|_2^2, \tag{16}$$

where

$$X_k = \langle x_k, e_1\rangle \cdot e_1 + \ldots + \langle x_k, e_i\rangle \cdot e_i + \ldots + \langle x_k, e_p\rangle \cdot e_p. \tag{17}$$

PCA under interval uncertainty is NP-hard: a conjecture. While the existing interval PCA algorithms are usually efficient, sometimes, they require a large amount of computation time. This empirical fact prompted a conjecture that the problem of PCA under interval uncertainty is NP-hard (see, e. g., [16, 23], also [19, 27] for formal definitions of NP-hardness).

This conjecture was also motivated by the fact that many similar statistical problems becomes NP-hard once we take interval uncertainty into account; even the problem of computing the range of the variance under interval uncertainty is NP-hard [9, 10, 20].

PCA under interval uncertainty is NP-hard: a proof. A part of the PCA problem is checking whether it is possible to achieve the *exact* data reduction, i. e., whether it is possible to find the vectors $e_1, \ldots, e_p$, $p < d$, and the values $x_{k,i}$ for which $X_k = x_k$ for all objects k.

In mathematical terms, this means checking whether it is possible to select the "column" vectors

$$z_i \overset{\text{def}}{=} (x_{1,i}, \ldots, x_{k,i}, \ldots, x_{n,i}) \tag{18}$$

in such a way that they are linearly dependent, i. e., that there exists a vector $\alpha = (\alpha_1, \ldots, \alpha_d) \neq 0$ for which for every k, we have

$$\alpha_1 \cdot x_{k,1} + \ldots + \alpha_i \cdot x_{k,i} + \ldots + \alpha_d \cdot x_{k,d} = 0. \tag{19}$$

For a square matrix ($d = n$), the existence of such α_i is equivalent to the matrix being singular. Thus, for a square interval matrix with entries $\mathbf{x}_{k,i}$, the possibility of such a reduction is equivalent to the possibility of finding a singular matrix with entries $x_{k,i} \in \mathbf{x}_{k,i}$. It is known that checking for the existence of such a matrix — or, equivalently, checking whether all matrices $x_{k,i} \in \mathbf{x}_{k,i}$ are non-singular — is NP-hard. This result — one of the first NP-hardness results in interval computations — was proved by S. Poljak and J. Rohn in [29] (see [19] for further similar results).

Thus, PCA under interval uncertainty is indeed NP-hard.

3. Realistic cases of interval-valued PCA and their computational complexity: formulation of the problem and the main result

The known NP-hardness result: reminder. The above result shows that, in general, the problem of PCA under interval uncertainty is NP-hard.

The general case is rare. The above proof is based on considering matrices in which all the entries are non-degenerate intervals.

In practice, however, often, most measurements are reasonably accurate, and only a few (or even one) variables are measured with significant uncertainty. In such situations, for each object k, we can safely assume that

— we know most of the values $x_{k,i}$ exactly, and

— only for a few i, we know the (non-degenerate) interval $\mathbf{x}_{k,i}$.

Natural question. When we consider such situations, will the interval-valued PCA still be NP-hard?

Simplest case. In particular, the same question about the computational complexity can be asked about the simplest case, when for each object k, at most one data point comes with interval uncertainty.

Our main result. Our result is that even for this simplest case, the data-reducing PCA problem under interval uncertainty is NP-hard.

Comment. Our proof will follow the main ideas from NP-hardness proofs described in [19].

4. Proof of the main result

What is NP-hard: a brief informal reminder. Crudely speaking, the fact that a problem $\mathcal{P}_0$ is NP-hard means that every problem $\mathcal{P}$ (from a reasonable class NP) can be reduced to this problem $\mathcal{P}_0$, i. e., informally, that this problem $\mathcal{P}_0$ is the toughest possible.

How NP-hardness is usually proved: by reduction to NP-hardness of a known problem. The usual way to prove NP-hardness of a problem $\mathcal{P}_0$ is to show that a known NP-hard problem $\mathcal{P}_k$ can be reduced to a particular case of our problem $\mathcal{P}_0$.

Since the problem $\mathcal{P}_k$ is NP-hard, this means that every problem $\mathcal{P}$ from the class NP can be reduced to this problem $\mathcal{P}_k$. Since the problem $\mathcal{P}_k$ can be, in turn, reduced to $\mathcal{P}_0$, this means that every problem $\mathcal{P}$ from the class NP can be reduced to $\mathcal{P}_0$. By definition of NP-hardness, this means that our problem $\mathcal{P}_0$ is indeed NP-hard.

Selection of the known NP-hard problem. As the known NP-hard problem $\mathcal{P}_k$, we take the following problem:

— we are given several positive integers $s_1 > 0, \ldots, s_m > 0$;

— we need to find the signs $\varepsilon_\ell \in \{-1, 1\}$ for which the corresponding signed sum of the given integers is equal to 0:

$$\sum_{\ell=1}^{m} \varepsilon_\ell \cdot s_\ell = 0. \tag{20}$$

Possible intuitive interpretation of the problem $\mathcal{P}_k$. The requirement (20) can be reformulated as

$$\sum_{\ell:\ \varepsilon_\ell=1} s_\ell - \sum_{\ell':\ \varepsilon_{\ell'}=-1} s_{\ell'} = 0. \tag{21}$$

If we move all the negative terms in the signed sum (21) into the other side, we get the equality between the sum of all the values to which we assigned plus and the sum of all the values to which we assigned minus:

$$\sum_{\ell:\ \varepsilon_\ell=1} s_\ell = \sum_{\ell':\ \varepsilon_{\ell'}=-1} s_{\ell'}. \tag{22}$$

The resulting problem allows the simple interpretation — e. g., as the problem of diving the inheritance into two equal parts:

— we have m objects with known costs s_i;

— we must divide them into two groups of equal cost.

Reduction: a reminder. To prove the NP-hardness of our interval-valued PCA problem $\mathcal{P}_0$, we wan to reduce the problem $\mathcal{P}_k$ to our problem $\mathcal{P}_0$.

To reduce means that for every instance $s_1, \ldots, s_m$ of the problem $\mathcal{P}_k$, we must form a case of the interval PCA problem $\mathcal{P}_k$ from whose solution we will be able to extract the solution to the original instance.

How we reduce. In the original problem, we have m positive integers $s_1, \ldots, s_m$ (and we must find the signs ε_i for which the signed sum is zero).

In the reduction, we form $n = 2m + 1$ objects with $d = m + 1$ variables and the following data:

— for the first m objects $\ell = 1, \ldots, m$, we take

$$\mathbf{x}_{\ell,\ell} = [-1, 1], \quad \mathbf{x}_{\ell,m+1} = [1, 1], \quad \mathbf{x}_{\ell,i} = [0, 0] \text{ for } i \neq \ell, m + 1; \tag{23}$$

— for the next m objects $k = m + \ell$, where $\ell = 1, \ldots, m$, we take

$$\mathbf{x}_{m+\ell,\ell} = [-1, 1], \quad \mathbf{x}_{m+\ell,m+1} = [1, 1],$$

$$\mathbf{x}_{m+\ell,i} = [0, 0] \text{ for } i \neq \ell, m + 1; \tag{24}$$

— finally, for the last object $k = 2m + 1$, we take

$$\mathbf{x}_{2m+1,\ell} = [s_\ell, s_\ell] \text{ for all } \ell \leq m, \quad \mathbf{x}_{2m+1,m+1} = [0, 0]. \tag{25}$$

Towards proving that this is indeed a reduction: what does it mean to have a solution to the instance of the interval PCA problem. In the interval PCA problem, we check whether the columns of the data matrix are linearly dependent, i. e., whether there exist values $x_{k,i}$ from the corresponding intervals $\mathbf{x}_{k,i}$ and values $\alpha = (\alpha_1, \ldots, \alpha_{m+1})$ not all equal to 0 for which the corresponding linear combination is equal to 0 for all objects k:

$$\alpha_1 \cdot x_{k,1} + \ldots + \alpha_i \cdot x_{k,i} + \ldots + \alpha_d \cdot x_{k,d} = 0. \tag{26}$$

Proving reduction: let us first consider the second group of m equations. For each $\ell \leq m$, from the $(m+\ell)$-th equation (24), we conclude that for some $x_{m+\ell,m+1} \in [-1,1]$, we get

$$\alpha_\ell + \alpha_{m+1} \cdot x_{m+\ell,m+1} = 0, \tag{27}$$

i. e.,

$$\alpha_\ell = -\alpha_{m+1} \cdot x_{m+\ell,m+1}. \tag{28}$$

Thus, the absolute value of α_ℓ is equal to the product of absolute values of a_{m+1} and of the absolute value of $x_{m+\ell,m+1}$:

$$|\alpha_\ell| = |\alpha_{m+1}| \cdot |x_{m+\ell,m+1}|. \tag{29}$$

Since $x_{m+\ell,m+1}$ is between -1 and 1, this means that the absolute value of α_ℓ is smaller than or equal to the absolute value of α_{m+1}:

$$|\alpha_\ell| \leq |\alpha_{m+1}|. \tag{30}$$

So, if the last α coefficient α_{m+1} is 0, then all the alpha values are zeros. Since we assumed that the vector α is not 0, this means that the coefficient α_{m+1} is not 0.

Since $\alpha_{m+1} \neq 0$, we can divide all the other alpha terms by this coefficient. For the resulting ratios

$$\varepsilon_\ell \stackrel{\text{def}}{=} \frac{\alpha_\ell}{\alpha_{m+1}}, \tag{31}$$

the inequality (30) implies that

$$|\varepsilon_\ell| \leq 1. \tag{32}$$

Proving reduction: let us now consider the first group of m equations. For each $\ell \leq m$, the ℓ-th equation implies that

$$x_{\ell,\ell} \cdot \alpha_\ell + \alpha_{m+1} = 0 \tag{33}$$

for some $x_{\ell,\ell} \in [-1,1]$, i. e., that

$$x_{\ell,\ell} \cdot \alpha_\ell = -\alpha_{m+1}. \tag{34}$$

Dividing both sides of this equality by $\alpha_{m+1} \neq 0$, we get

$$x_{\ell,\ell} \cdot \varepsilon_\ell = -1. \tag{35}$$

So, the product of the absolute values of ε_ℓ and of $x_{\ell,\ell}$ is 1:

$$|x_{\ell,\ell}| \cdot |\varepsilon_\ell| = 1, \tag{36}$$

and

$$|\varepsilon_\ell| = \frac{1}{|x_{\ell,\ell}|}. \tag{37}$$

Since $x_{\ell,\ell}$ is between -1 and 1, its absolute value is bounded by 1. Hence, the absolute value of ε_ℓ is at least one:

$$|\varepsilon_\ell| \geq 1. \tag{38}$$

From (38) and (32), we conclude that $|\varepsilon_\ell| = 1$, i. e., that $\varepsilon_\ell \in \{-1, 1\}$.

Proving reduction: let us use the last equation. The last equation has the form

$$\sum_\ell \alpha_\ell \cdot s_\ell = 0. \tag{39}$$

Dividing both sides by α_{m+1}, we get

$$\sum_{\ell} \varepsilon_{\ell} \cdot s_{\ell} = 0, \tag{40}$$

which is exactly what we wanted.

The reduction is proven. Thus, the PCA under interval uncertainty problem is indeed NP-hard even in the simplest case when for each object, no more than one quantity is known with interval uncertainty.

Acknowledgments

The author is thankful to all his colleagues from the Human Neuroimaging Lab, specially to Drs. P. Read Montague and Terry Lohrenz, for valuable suggestions, and to the anonymous referees for their help.

References

[1] Antoch J., Brzezina M., Miele R. A note on variability of interval data // Comput. Statist. 2010. Vol. 25. P. 143–153.

[2] Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data / Eds. H.H. Bock and E. Diday. Heidelberg: Springer-Verlag, 2000.

[3] Buxton R.B. An Introduction to Functional Magnetic Resonance Imaging. Principies and Techniques. Cambridge, Massachusetts: Cambridge Univ. Press, 2002.

[4] Handbook of Functional Neuroimaging of Cognition / Eds. R. Cabeza and A. Kingstone. Cambridge, Massachusetts: MIT Press, 2006.

[5] Cazes P., Chouakria A., Diday E., Schektman Y. Extension de l'analyse en composantes principales à des données de type intervalle // Revue de Statist. Appl. 1997. Vol. 45. P. 5–24.

[6] Chouakria A. Extension de L'analyse en Composantes Principales á des Données de Type Intervalle. PhD Dissertation. Univ. of Paris IX Dauphine, 1998.

[7] Chouakria A., Diday E., Cazes P. An improved factorial representation of symbolic objects // Proc. of the Conf. of Knowledge Extraction and Symbolic Data Analysis KESDA'98. Luxembourg, 1999.

[8] D'Urso P., Giordani P. A least squares approach to principal component analysis for interval valued data // Chemometr. and Intelligent Laboratory Systems. 2004. Vol. 70, No. 2. P. 179–192.

[9] Ferson S., Ginzburg L., Kreinovich V. et al. Computing variance for interval data is NP-hard // ACM SIGACT News. 2002. Vol. 33, No. 2. P. 108–118.

[10] Ferson S., Ginzburg L., Kreinovich V. et al. Exact bounds on finite populations of interval data // Reliable Comput. 2005. Vol. 11, No. 3. P. 207–233.

[11] Ferson S., Kreinovich V., Hajagos J. et al. Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Sandia National Laboratories. Rep. SAND2007-0939, May 2007.

[12] Gioia F. Statistical Methods for Interval Variables. PhD Dissertation. Department of Mathematics and Statistics. Univ. of Naples "Federico II", 2001 (in Italian).

[13] Gioia F., Lauro C.N. Principal component analysis on interval data // Comput. Statist. 2006. Vol. 21. P. 343–363.

[14] Hu C., Kearfott R.B. Interval matrices in knowledge discovery // Knowledge Processing with Interval and Soft Computing / Eds. C. Hu, R.B. Kearfott, A. de Korvin, and V. Kreinovich. London: Springer-Verlag, 2008. P. 99–117.

[15] Huettel S.A., Song A.W., McCarthy G. Functional Magnetic Resonance Imaging. Sinauer Associates. Sunderland, Massachusetts, 2004.

[16] Irpino A. Spaghetti' PCA analysis: An extension of principal components analysis to time dependent interval data // Patt. Recognit. Lett. 2006. Vol. 27. P. 504–513.

[17] Functional MRI: An Introduction to Methods / Eds. P. Jezzard, P.M. Matthews, and S.M. Smith. New York: Oxford Univ. Press, 2003.

[18] Jolliffe I.T. Principal Component Analysis. New York: Springer-Verlag, 2002.

[19] Kreinovich V., Lakeyev A., Rohn J., Kahl P. Computational Complexity and Feasibility of Data Processing and Interval Computations. Dordrecht: Kluwer, 1998.

[20] Kreinovich V., Xiang G., Starks S.A. et al. Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity // Reliable Comput. 2006. Vol. 12, No. 6. P. 471–501.

[21] Lauro C., Palumbo F. Principal component analysis of interval data: A symbolic data analysis approach // Comput. Statist. 2000. Vol. 15. P. 73–87.

[22] Lauro C., Palumbo F. Some results and new perspectives in principal component analysis for interval data // Atti del Convegno CLADAG'03 Gruppo di Classificazione della Società Italiana di Statistica, 2003. P. 237–244.

[23] Lauro C.N., Verde R., Irpino A. Principal component analysis of symbolic data described by intervals // Symbolic Data Analysis and the SODAS Software / Eds. E. Diday and M. Noirhome Fraiture. Chichester, UK: John Wiley and Sons, 2007. P. 279–312.

[24] Lauro C., Palumbo F. Principal component analysis for non-precise data // New Developments in Classification and Data Analysis / Eds. M. Vichi, P. Monari, S. Mignani and A. Montanari. Berlin, Heidelberg, New York: Springer-Verlag, 2005. P. 173–184.

[25] Moore R.E., Kearfott R.B., Cloud M.J. Introduction to Interval Analysis. Philadelphia, Pennsylvania: SIAM Press, 2009.

[26] Palumbo F., Lauro C. A PCA for interval-valued data based on midpoints and radii // New Developments on Psychometrics: Proc. of the Intern. Meeting of the Psychometric Society IMPS'2001 / Eds. H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J.J. Meulman. Tokyo: Springer-Verlag, 2003. P. 641–648.

[27] Papadimitriou C.H. Computational Complexity. Addison Wesley, 1994.

[28] Pearson K. On lines and planes of closest fit to systems of points in space // Philosop. Magazine. 1901. Vol. 2, No. 6. P. 559–572; available as `http://stat.smmu.edu.cn/history/pearson1901.pdf`

[29] Poljak S., Rohn J. Checking robust nonsingularity is NP-hard // Math. Control Signals Syst. 1993. Vol. 6, No. 1. P. 1–9.

[30] Rodriguez O. Classification et Modeles Lineaires en Analyse des Donnes Symboliques. PhD Dissertation. Univ. of Paris IX Dauphine, 2000.

[31] Sato-Ilic M. Weighted principal component analysis for interval-valued data based on fuzzy clustering // Proc. of the 2003 IEEE Conf. on Systems, Man, and Cybernetics. IEEE Press, 2003. P. 4476–4482.

Received for publication 29 March 2010