ISSN 1392-8619 print/ISSN 1822-3613 online
ÛKIO TECHNOLOGINIS IR EKONOMINIS VYSTYMAS
TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY
http://www.tede.vgtu.lt
2006, Vol XII, No 1, 69–75

ON MULTIDIMENSIONAL SCALING WITH EUCLIDEAN AND CITY BLOCK METRICS

Antanas Žilinskas[1], Julius Žilinskas[2]

Institute of Mathematics and Informatics, Akademijos g. 4, LT-08663 Vilnius, Lithuania
E-mail: [1]antanasz@ktl.mii.lt, [2]julius.zilinskas@mii.lt

Received 3 October 2005; accepted 4 January 2006

Abstract. Experimental sciences collect large amounts of data. Different techniques are available for information elicitation from data. Frequently statistical analysis should be combined with the experience and intuition of researchers. Human heuristic abilities are developed and oriented to patterns in space of dimensionality up to 3. Multidimensional scaling (MDS) addresses the problem how objects represented by proximity data can be represented by points in low dimensional space. MDS methods are implemented as the optimization of a stress function measuring fit of the proximity data by the distances between the respective points. Since the optimization problem is multimodal, a global optimization method should be used. In the present paper a combination of an evolutionary metaheuristic algorithm with a local search algorithm is used. The experimental results show the influence of metrics defining distances in the considered spaces on the results of multidimensional scaling. Data sets with known and unknown structure and different dimensionality (up to 512 variables) have been visualized.

Keywords: Multidimensional scaling, global optimization, metaheuristics, city block metrics, visualization of multidimensional data.

1. Introduction

Multidimensional scaling (MDS) is a technique for the analysis of multidimensional data widely usable in different applications [1]. The dissimilarity between pairs of n objects is given by matrix δ_{ij}, $i, j = 1, \ldots, n$, and it is supposed that $\delta_{ij} = \delta_{ji}$. Points x_i, $i = 1, \ldots, n$ in embedding space $\mathbf{R}^m$ should be found which interpoint distances fit given dissimilarities. The problem is reduced to minimization of a fitness criterion, e.g. so called *STRESS* function

$$S(X) = \sum_{i<j} w_{ij}(d_{ij}(X) - \delta_{ij})^2 ,$$

where $X = (x_{11}, \ldots, x_{n1}, x_{12}, \ldots, x_{nm})$, and $d_{ij}(X)$ denotes the distance between points x_i and x_j. It is supposed that the weights are positive: $w_{ij} > 0$, $i, j = 1, \ldots, n$. Most frequently two dimensional ($m = 2$) embedding vector space $\mathbf{R}^m$ is considered, for example, aiming to visualize the results of MDS.

Although *STRESS* function is defined by the analytical formula which seems rather simple, its minimization is difficult. *STRESS* function normally has many local minima. The minimization problem is high dimensional: $X \in \mathbf{R}^N$ where the number of variables is equal to $N = n \times m$. Smoothness of *STRESS* depends on metrics of embedding space, however, nondifferentiability normally can not be ignored. Therefore MDS is a difficult global optimization problem. Global optimization methods are developed for various classes of multimodal problems [2]. Different global optimization methods have been applied to MDS, e.g. tunneling method in [3], evolutionary method in [4], simulated annealing in [5–7].

2. Comparison of two metrics in the embedding space

Majority of publications on MDS consider *STRESS* with Euclidean distances:

$$d_{ij}^{ED}(X) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} .$$

However, the interest to the methods based on city block distances increases:

$$d_{ij}^{CB}(X) = \sum_{k=1}^{m} |x_{ik} - x_{jk}|;$$

see, e.g. [5, 6]. For a review on MDS with city block distances we refer to [5]. Points x_i, defined using different distances in the embedding space, can be interpreted as different projections of the original objects in the embedding space. Different projections can provide different information on original objects thus enhancing exploratory power of MDS methods. To understand what properties of the original objects can be best highlighted using Euclidean and city block metrics in the present paper we investigate the images of the objects with known properties.

For this experiment a reliable global optimization method is desirable. The results in [8–9] show that a combination of evolutionary global search with an efficient local descent is the most time consuming, but the most reliable method. The implementation suitable for both versions of *STRESS* function is favorable. Therefore we have implemented a memetic algorithm with convex crossover operator and uniform selection. To avoid difficulties caused by nondifferentiability for local minimization a search algorithm is preferable against a gradient based descent algorithm. However, a good convergence rate of the algorithm for smooth functions is desirable. The known algorithm by Powell has the properties mentioned above; the implementation of the algorithm from [10] has been chosen.

We use several sets of data for testing. There are many applications of MDS where data (objects) are a set of multidimensional points. Dissimilarity is defined by the distance in an original vector space. Such a problem is important, e.g. for the analysis of bio-medical data. A classical example is Iris data [11]. For an example of visualization of objective and subjective data of the population of patients we refer to [12].

Iris data are analyzed using different methods and they have a well known structure: there are two contiguous clusters and one well separated cluster. Images of Iris data obtained by means of different MDS methods are expected to highlight this structure of the data. There are 150 instances (50 in each of three classes: Iris Setosa, Iris Versicolour, Iris Virginica) in this data. 4 numeric attributes (sepal length, sepal width, petal length, petal width) define multidimensional data. Therefore, dim = 4, n = 150.

The other data sets represent well understood geometric objects: vertexes of multidimensional cubes and simplexes of different dimensionality. For both types of objects symmetric location of vertexes is characteristic. In the image of a simplex special central location of the "zero" vertex is expected while the other vertexes are expected to be shown alike. All vertexes of a hypercube are equally far from the center and compose clusters containing 2^d points.

Vertexes of multidimensional simplex may be defined by

$$v_{ij} = \begin{cases} 1, & \text{if } i = j+1, \\ 0, & \text{otherwise,} \end{cases}$$

where i = 1, …, dim+1, and j = 1, …, dim. The number of vertexes of multidimensional simplex is n = dim+1.
The number of vertexes of multidimensional hypercube is $n = 2^{\dim}$.

Finally a data set of a psychological experiment related to soft drinks testing is used. In this case considered objects are defined by means of a dissimilarity matrix where dissimilarities are measured experimentally [13]. There are ten objects representing each soft drink, therefore n = 10.

3. Experimental Results

The developed evolutionary global search algorithm with different metrics has been used to visualize data sets. The normalized best function value found is shown in figures:

$$f(X) = \sqrt{\frac{2S(X)}{n \times (n-1)}},$$

weights are supposed to be w_{ij}=1.

The images of Iris data visualized using multidimensional scaling are shown in Fig 1. Four pictures show the comparison of different metrics. In the original space city block metric is used in the upper row and Euclidean in the lower row. In the embedding space city block metric is used in the left column and Euclidean in the right column. Different classes of Iris data are denoted by letters t (Iris Setosa), l (Iris Versicolour) and g (Iris Virginica). The known structure of the data is well visible in all pictures. However, the contiguous clusters are best separated in case Euclidean metric is used to measure distances in original space, and city block metric is used in embedding space.

The images of simplexes and hyper-cubes visualized using multidimensional scaling are shown in Fig 2 and Fig 3. Four columns show the comparison of different metrics. The first column represents visualization when both original and embedding spaces are with city block metric, the secon – with city block original space and Euclidean embedding space, the third – with Euclidean original and city block embedding spaces, and the fourth – with both Euclidean spaces.

The vertexes of the objects are visualized and are shown as circles. To make representations more visual, adjacent vertices are joined by lines. Darker lines show joins adjacent to the "zero" vertex in the case of simplex and adjacent to two opposite vertexes in the case of hyper-cube.

Visualized three-dimensional to twenty-dimensional simplexes are shown in Fig 2. When city block metric is used in both spaces, three- and four-dimensional simplexes can be visualized fitting multidimensional distances exactly. This is the case for city block metrics only. The "zero" vertex is always visualized in the center of the structure. Other vertexes of multidimensional simplexes tend to form a rhomb or diamond shaped structure when city block metric is used in the embedding space. Let us note that points on the rhomb are of the same distance to the center when city block metric is used. In the case of Euclidean embedding space, other vertexes of multidimensional simplexes tend to form a circle, however, when the dimensionality of simplex and the number of vertexes are increased, a part of the vertexes starts to form a smaller circle. In this case "non-zero" vertexes are not shown alike.

Visualized three-dimensional to eight-dimensional hyper-cubes are shown in Fig 3. Perspective is usually used by artists and designers to visualize the cube on the plain. It would be possible to imagine the cube visualized by MDS with city block metrics, if positive perspective is used in one coordinate direction and negative in the other. Similarly to simplexes, vertexes of multidimensional hyper-cubes tend to form a diamond shaped structure when city block metric is used in the embedding space. However, in the case of hyper-cubes, visualized vertexes form clusters. In the case of Euclidean embedding space, vertexes of hyper-cube tend to form clusters and fill a circle. In this case vertexes are not shown alike again.

Diamond shaped structures in visualization with city block embedding space suggest to use modified city block metric rotated by 45 degrees which can be called diagonal

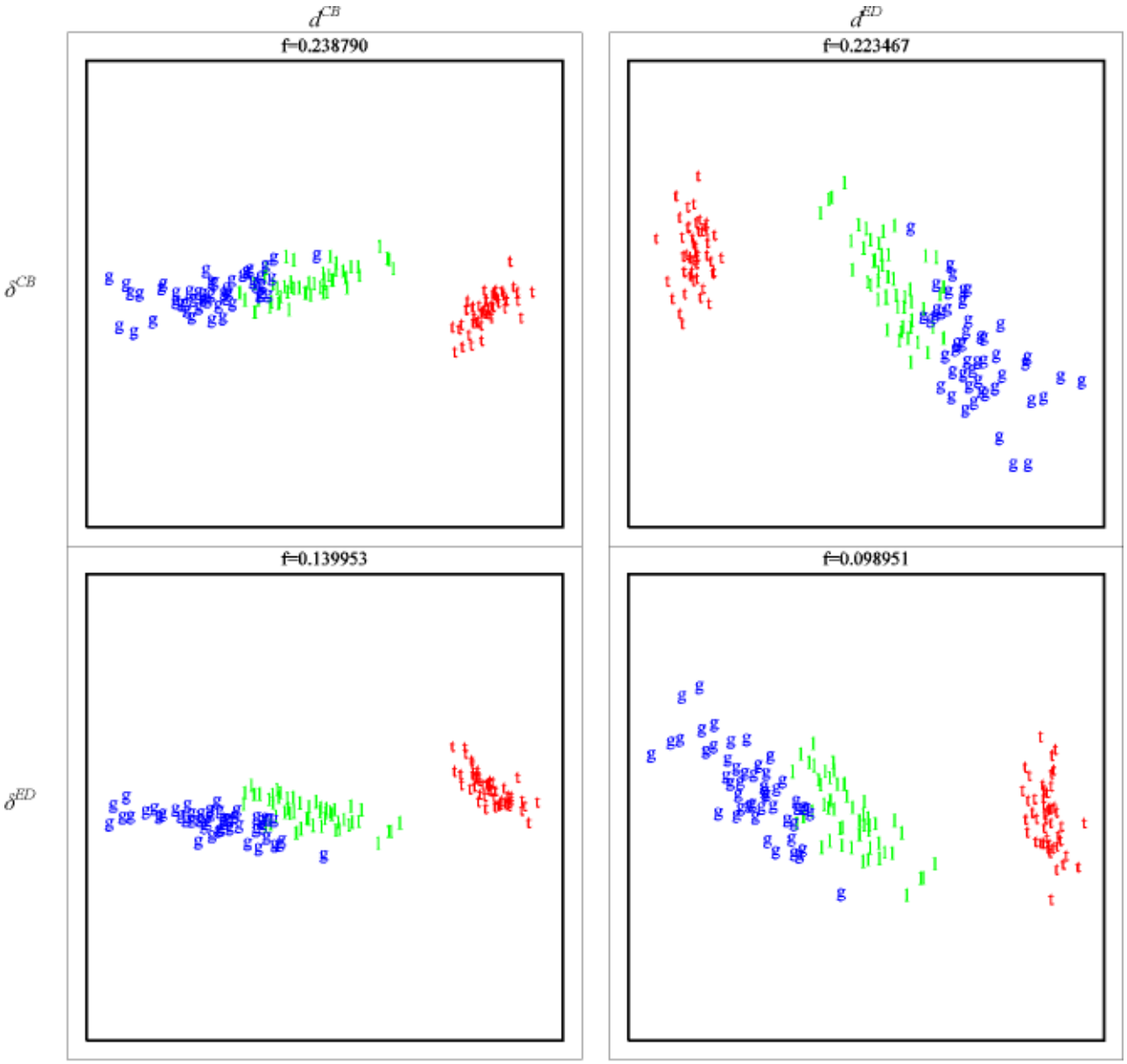

Fig 1. Images of Iris data

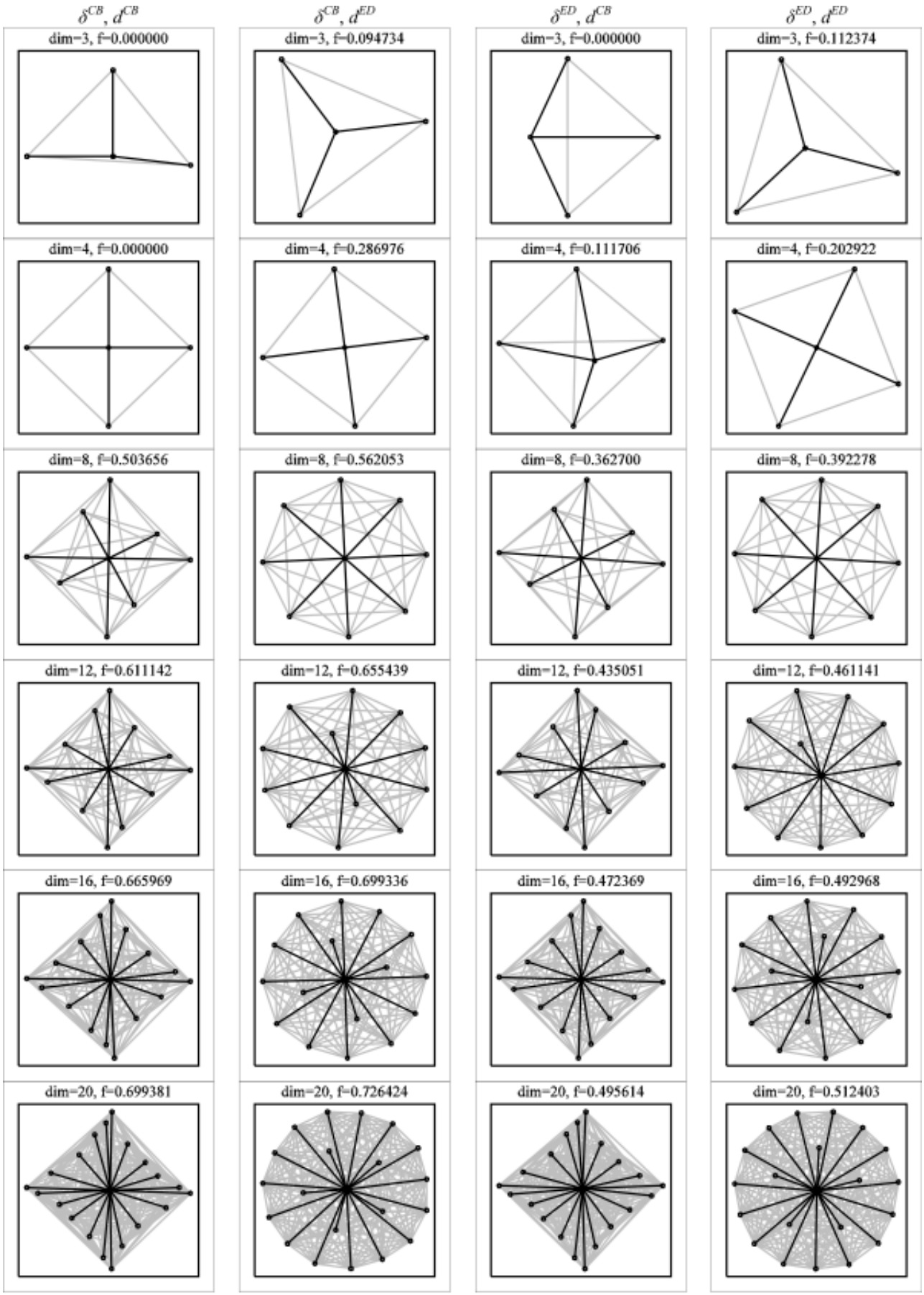

Fig 2. Images of simplexes

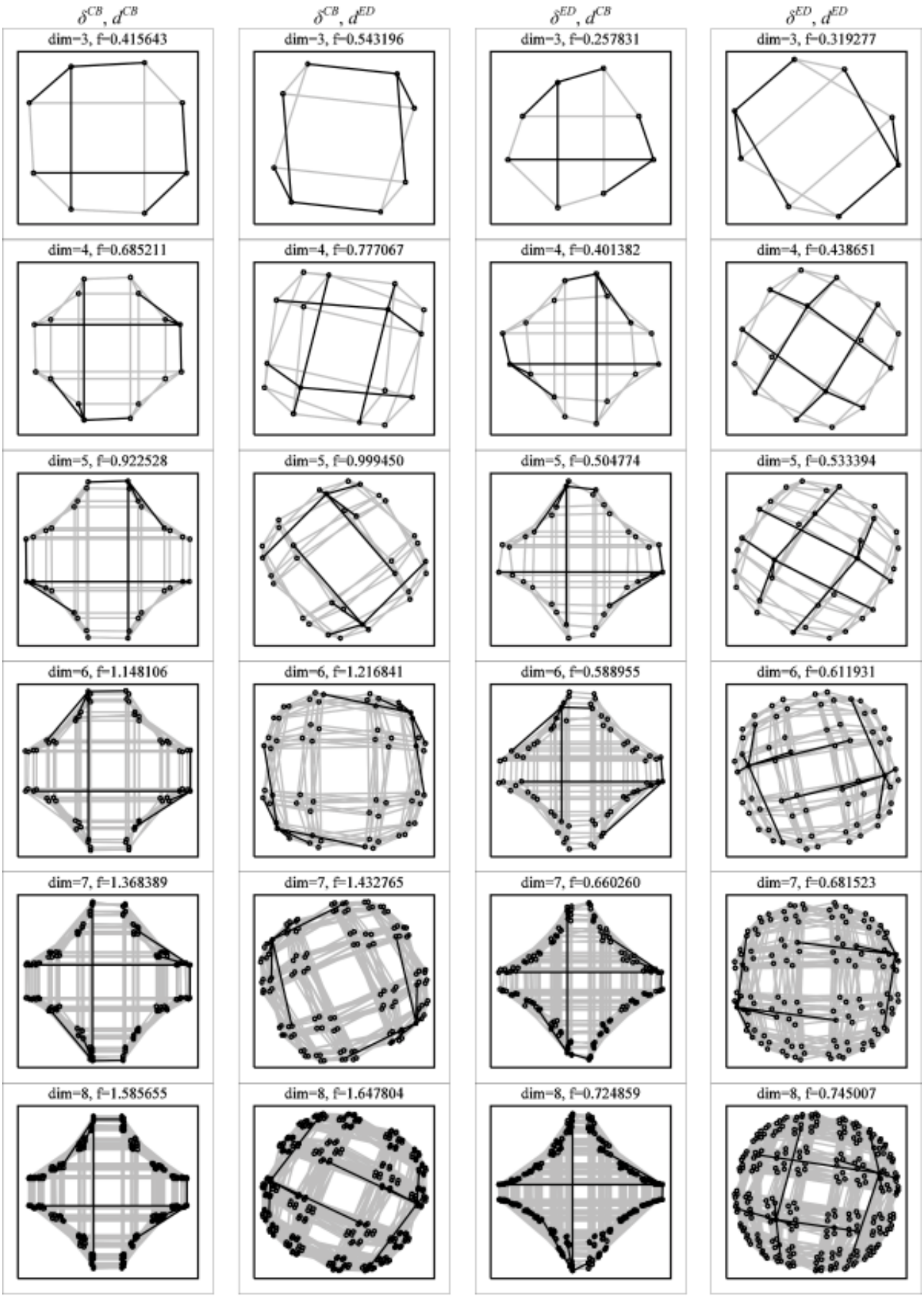

Fig 3. Images of hyper-cubes

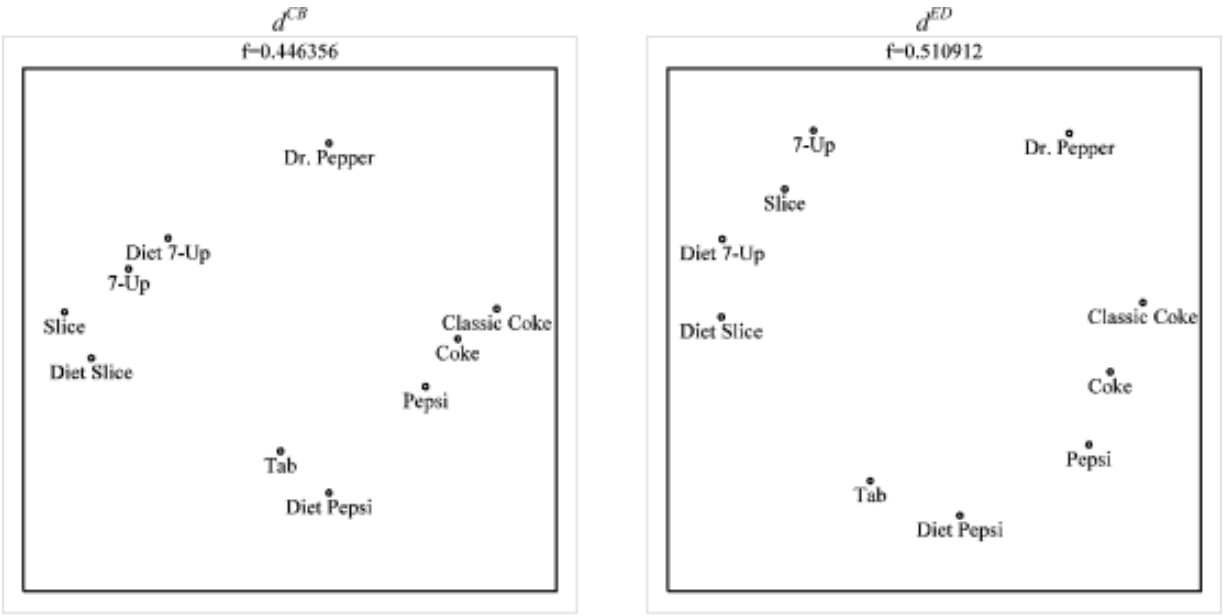

Fig 4. Images of visualized soft drinks

city block metric, serpentine metric or staircase metric. Multidimensional simplexes and hyper-cubes visualized with such metric would form structures similar to squares.

The images of soft drinks visualized using multidimensional scaling are shown in Fig 4. Visualization with city block embedding space is shown on the left and with Euclidean embedding space – on the right. Objects tend to form a diamond shaped structure when city block metric is used and an ellipse shaped structure when Euclidean metric is used. However, both pictures are similar and can be interpreted similarly; therefore it is difficult to assess the advantages of metrics.

4. Conclusions

Visualization results depend on metric of the embedding space stronger than on metric of the space of originals.

Visualization of geometric data better highlights the properties of the originals when distances in embedding space are measured according to city block metric than according to Euclidean metric.

For the data without strictly defined structure visualization the use of both investigated metrics gives similar pictures; in this case it is difficult to assess the advantages of visualization methods.

5. Acknowledgements

The work of the second co-author is supported by the NATO Reintegration grant CBP.EAP.RIG.981300.

References

1. Borg, I. and Groenen, P. Modern Multidimensional Scaling. NY: Springer, 1997.
2. Törn, A. and Žilinskas, A. Global Optimization, Berlin: Springer, 1989.
3. Groenen, P. The Majorization Approach to Multidimentional Scaling. Amsterdam: DSWO, 1993. 110 p.
4. Mathar, R. and Žilinskas, A. On global optimization in two-dimensional scaling. *Acta Applicandae Mathematicae*, Vol 33, 1993, p. 109–118.
5. Brusco, M. J. A simulated annealing heuristics for unidimensional and multidimensional (city block) scaling of symmetric proximity matrices. *Journal of Classification*, Vol 18, 2001, p. 3–33.
6. Leng, P. L. and Lau, K. Estimating the city-block two-dimensional scaling model with simulated annealing. *European Journal of Operational Research*, Vol 158, 2004, p. 518–524.
7. Klock, H. and Buhman, J. Data visualization by multidimensional scaling: a deterministic annealing approach. *Pattern Recognition*, Vol 33, No 4, 1999, p. 651–669.
8. Mathar, R. A hybrid global optimization algorithm for multidimensional scaling. In: Classification and Knowledge Organization, Eds. R.Klar and O. Opitz. Heidelberg: Springer, 1997, p. 63–71.
9. Groenen, P. J. F.; Mathar, R. and Trejos, J. Global optimization methods for multidimensional scaling applied to mobile communications. In: W. Gaul, O. Opitz, M. Schader (Eds.) Data Analysis: Scientific Modelling and Practical Application. Berlin: Springer, 2000, p. 459–470.
10. Press, W. et al. Numerical Recipes in C++. Cambridge: Cambridge University Press, 2002.
11. Fisher, R. A. The use of multiple measurements in taxonomy problems. *Annals of Eugenics*, Vol 7, 1936, p. 179–188.
12. Podlipskytė, A.; Žilinskas, A.; Žemaitytė, D. and Varoneckas, G. A new version of MDS method and its application for visualization of data on sleep quality. In: Pattern Recognition and Information Processing: PRIP'2001, sixth international conference, Minsk, Vol 2, 2001, p. 181–188.
13. Green, P. E.; Carmone, Jr. F. J.; and Smith, S. M. Multidimensional Scaling: Concepts and Applications. Boston: Allyn and Bacon, 1989.

DAUGIAMAČIŲ SKALIŲ SU EUKLIDO IR MANHETENO METRIKOMIS SUDARYMO METODAI

A. Žilinskas, J. Žilinskas

Santrauka

Eksperimentiniai mokslai kaupia didelius duomenų kiekius. Sukurta daug metodų informacijai iš duomenų išgauti. Dažnai statistiniai metodai yra derinami su euristine analize pagrįsta tyrinėtojų intuicija. Tačiau euristiniai žmonių sugebėjimai gerai tinka analizuoti duomenis, kurių matavimų skaičius neviršija 3. Daugiamačių skalių metodas skirtas vaizduoti objektams mažo matavimų skaičiaus erdvėje, kai objektai apibrėžti panašumais/nepanašumais, o atstumai vaizdų erdvėje vaizduoja nepanašumus. Daugiamačių skalių metodai sudaromi kaip vaizdavimo tikslumo kriterijaus, paprastai vadinamo stresu, minimizavimo procedūros. Kadangi optimizavimo uždaviniai daugiaekstremalūs, jiems spręsti reikia globalios optimizacijos metodų. Šiame darbe pasiūlytas algoritmas, jungiantis metaeuristinę globalią paiešką ir lokalios minimizacijos metodą. Eksperimentais ištirta metrikos vaizdų erdvėje įtaka vaizdavimo tikslumui ir algoritmo efektyvumui. Eksperimentuose naudotos duomenų aibės su žinoma ir nežinoma struktūra; optimizacijos uždavinio kintamųjų yra iki 512.

Pagrindiniai žodžiai: daugiadimensės skalės, globalioji optimizacija, metaeuristiniai metodai, Manheteno metrika, daugiamačių duomenų vizualizacija.

Antanas ŽILINSKAS. Professor and head of Department of Applied Informatics Institute of Mathematics and Informatics. Doctor of Science (technical cybernetics) (1973) Kaunas University of Technology, Doctor of Mathematical Science (Habilitation, 1985). St. Petersburg University, Lithuanian National Award for scientific achievements of 2001. Research interests: statistical global optimization theory, algorithms and applications, more than 100 publications including 5 monographs and 6 textbooks. Memberships: IEEE including CS and CIS, International Engineering Academy, American Mathematical Society, IFIP WG 7.6. Editorial boards: Journal of Global Optimization, Control and Cybernetics, Informatica**.**

Julius ŽILINSKAS. Doctor of Science, senior researcher, Systems Analysis Department, Institute of Mathematics and Informatics. Doctor of Science (informatics engineering) (2002), Kaunas University of Technology. Studies at City University of London, Technical University of Denmark, University of Copenhagen, Edinburgh Parallel Computing Centre. Postdoctoral research follow at University College of London (UK, 12 months). Over 20 publications in scientific journals and proceedings. Research interests: global optimization, data analysis and parallel computing.