



ON VISUALIZATION OF MULTIDIMENSIONAL DATA USING THREE-DIMENSIONAL EMBEDDING SPACE

Antanas Žilinskas¹, Julius Žilinskas²

Institute of Mathematics and Informatics, Akademijos g. 4, LT-08663 Vilnius, Lithuania

E-mail: ¹antanasz@ktl.mii.lt; ²julius.zilinskas@mii.lt

Received 15 June 2006; accepted 20 November 2006

Abstract. Multidimensional scaling addresses the problem of representation of objects specified by proximity data by points in low dimensional embedding space. The problem is reduced to optimization of an accuracy measure of fit of the proximity data by the distances between the respective points. Three-dimensional embedding space is considered in the present paper. Images of data of different dimensionality are discussed as well as dependence of visualization accuracy on dimensionality of embedding space and complexity of data.

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

1. Introduction

Multidimensional scaling (MDS) is a technique for exploratory analysis of multidimensional data widely usable in different applications [1]; e.g. applications of MDS based visualization of observation points in interactive optimization systems is discussed in [2]. Let us give a short formulation of the problem. Pairwise dissimilarities between n objects are given by the matrix (δ_{ij}) , $i, j = 1, \dots, n$. The points in an m -dimensional embedding space $x_i \in \mathfrak{R}^m$, $i = 1, \dots, n$ should be found whose inter-point distances fit the given dissimilarities. Different measures of accuracy of fit can be chosen defining different images of the considered set of objects. In the case the objects are points in a high dimensional vector space such images can be interpreted as different nonlinear projections of the set of points in high dimensional space to an embedding space of lower dimensionality. The problem of construction of images of the considered objects is reduced to minimization of an accuracy of fit criterion, e.g. of the most frequently used least squares STRESS function:

$$S(X) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} (d(x_i, x_j) - \delta_{ij})^2,$$

where $X = (x_{11}, x_{21}, \dots, x_{nm})$; $d(x_i, x_j)$ denotes the distance between the points x_i and x_j ; it is supposed that the weights are positive: $w_{ij} > 0$, $i, j = 1, \dots, n$.

To define a particular criterion, a norm in \mathfrak{R}^m should be chosen implying the particular formula for calculating distances $d(x_i, x_j)$. In this paper city block distances are used:

$$d(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}|.$$

STRESS normally has many local minima. In the case of city block distances STRESS can be non differentiable even at the minimum point [3]. Therefore MDS with city block distances is a difficult high dimensional ($X \in \mathfrak{R}^N$, $N = n \times m$) global optimization problem.

2. Two level minimization in multidimensional scaling

Two level approach for minimization of STRESS with city block distances in two-dimensional embedding space is proposed in [3]. In the present paper a generalization of the algorithm of [3] from two-dimensional scaling to general case is considered. The piecewise quadratic structure of STRESS with city block distances is exploited suggesting a two level minimization algorithm: to solve a combinatorial problem at upper level, and to solve a quadratic programming problem at lower level:

$$\min_P S(P), \quad S(P) = \min_{X \in A(P)} S(X).$$

The upper level combinatorial problem is defined over the set of m permutations of $1, \dots, n$. Properties of the objective function $S(P)$ are not known, therefore optimal solution can not be found with a guarantee by means of an efficient algorithm. The lower level quadratic problem whose feasible region is defined by the sequence of coordinates $x_{1i}, x_{2i}, \dots, x_{ni}$, can be solved using a standard quadratic programming algorithm.

3. Experimental investigation

Minimum values of STRESS depend on data, e.g. the influence of number of objects is obvious from the formula. Such a dependency hinders the investigation of visualization accuracy on dimensionality of original and embedding spaces. To be able to compare accuracy of visualization of different data we use a relative error:

$$f(X) = \sqrt{\frac{S(X)}{\sum_{i=1}^n \sum_{j=1}^n w_{ij} \delta_{ij}^2}},$$

where impact of number of objects is reduced as well as scale of the proximity measure.

Performance of deterministic global optimization algorithms normally is measured using time of optimization and the smallest function value found. Performance of stochastic global optimization algorithms is measured similarly; in multidimensional scaling and in visualization of multidimensional data we take into account the best estimate of global minimum f^* found in 100 runs of 10 seconds each, and the percentage of runs (perc) where estimate of global, f^* exceeds no more than 10^{-4} .

3.1. Test problems

Several sets of multidimensional points corresponding to well understood geometric objects are needed for the experimental investigation. We want to choose difficult test problems, i.e. difficult to visualize geometric objects. The data with desired properties correspond to the multidimensional objects equally extending in all dimensions of the original space, e.g. sets of vertices of multidimensional cubes and simplexes. Dissimilarity between vertices is measured by the distance in the original vector space defined by its metric, city block metrics in our case. Global optimization problems of different difficulty can be constructed by defining dimensionality of the original space **dim**. Below we use shorthand ‘cube’ and ‘simplex’ for sets of their vertices.

The number of vertices of multidimensional cube is $n = 2^{\text{dim}}$, and the dimensionality of global minimization problem is $N = m \times 2^{\text{dim}}$. The coordinates of i -th vertex of a **dim**-dimensional cube are equal either to 0 or to 1, and

they are defined by binary code of $i = 1, \dots, n$.

The number of vertices of multidimensional simplex is $n = \text{dim} + 1$, and the dimensionality of global minimization problem is $N = m \times (\text{dim} + 1)$. Two types of multidimensional simplexes are used. The distances between any two vertices of equidistant simplex are equal: $\delta_{ij} = 1, i \neq j$. Vertices of right-angle simplex can be defined by:

$$v_{ij} = \begin{cases} 1, & i = j + 1 \\ 0, & \text{otherwise} \end{cases} \quad i = 1, \dots, \text{dim} + 1; \quad j = 1, \dots, \text{dim}.$$

3.2. On explicit enumeration of all feasible solutions of upper level combinatorial problem

The upper level combinatorial problem is defined over the set of m permutations of $1, \dots, n$. The number of feasible solutions is $(n!)^m$. The number of feasible solutions can be reduced taking into account symmetries of multidimensional scaling. Refusing mirrored solutions with respect to each coordinate the number of feasible solutions can be reduced to $((n/2)!)^m$ and refusing exchange of coordinate axes it can be reduced even further. Time in seconds (t) of explicit enumeration of feasible solutions considering symmetries, number of quadratic programming problems solved (NQPP), and values of relative errors at optimal solutions (f^*) are shown in Table 1. As expected, relative error grows when dimensionality of original space (and the number n of objects representing vertices) grows. Relative error reduces when dimensionality of embedding space m grows. As the number of feasible solutions grows very fast, the time of explicit enumeration of all solutions grows very fast too. As the time of explicit enumeration is not equal even when NQPP is equal, time to solve lower level quadratic programming problems varies.

3.3. Images

Evaluation of all feasible solutions of upper level combinatorial problem is computationally infeasible for all but the smallest problems. Properties of the objective function are not known, therefore optimal solution cannot be found by means of an efficient algorithm with a guarantee. As it was shown in [4, 5] the hybrid algorithm [6] combining evolutionary global search with efficient local descent is the most reliable though the most time consuming method for MDS with Euclidean distances. Similarly for MDS with city block metrics, a two level algorithm with evolutionary search at upper level seems prospective. Such an algorithm has been used in this paper to visualize multidimensional data in three-dimensional space.

Three-dimensional images can be shown using stereo screens. In this paper three-dimensional images are shown using orthogonal and isometric projections. Four projections are shown for each image: the left upper picture shows the orthogonal projection on the xz -plane, the right upper

picture shows the orthogonal projection on the yz-plane, the left lower picture shows the orthogonal projection on the xy-plane, and the right lower picture shows isometric projection. In the orthogonal projections the x-axis directs left, the z-axis directs up, and the y-axis directs right in the right upper picture and down in the left lower picture. In

the isometric projection the x-axis directs down left, the y-axis directs down right, and the z-axis directs up.

Projections of three-dimensional images of vertices of four-dimensional and five-dimensional cubes are shown in Fig 1. The vertices are visualized as circles. To make representations more visual, adjacent vertices are joined by lines.

Table 1. Explicit enumeration of all feasible solutions of upper level problem with considering symmetries

<i>n</i>	<i>m</i> = 1		<i>m</i> = 2		<i>m</i> = 3	
	<i>t, s</i> (NQPP)	<i>f</i> *	<i>t, s</i> (NQPP)	<i>f</i> *	<i>t, s</i> (NQPP)	<i>f</i> *
Cubes						
4	0.00 (12)	0.4082	0.00 (78)	0.0002	0.02 (364)	0.0002
8	0.24 (20160)	0.4787	12572.00 (203222880)	0.2245		
Equidistant simplexes						
3	0.00 (3)	0.3333	0.00 (6)	0.0002	0.00 (10)	0.0001
4	0.00 (12)	0.4082	0.00 (78)	0.0002	0.01 (364)	0.0001
5	0.00 (60)	0.4472	0.03 (1830)	0.1907	1.79 (37820)	0.0001
6	0.00 (360)	0.4714	1.71 (64980)	0.2309	589.72 (7840920)	0.0001
7	0.03 (2520)	0.4879	118.59 (3176460)	0.2621		
8	0.21 (20160)	0.5000	10229.00 (203222880)	0.2825		
9	2.24 (181440)	0.5092				
10	26.63 (1814400)	0.5164				
11	351.09 (19958400)	0.5222				
12	4702.00 (239500800)	0.5270				
Right-angle simplexes						
3	0.00 (3)	0.0003	0.00 (6)	0.0002	0.00 (10)	0.0001
4	0.00 (12)	0.3651	0.00 (78)	0.0002	0.01 (364)	0.0001
5	0.00 (60)	0.4140	0.04 (1830)	0.0002	2.02 (37820)	0.0001
6	0.01 (360)	0.4554	2.05 (64980)	0.1869	661.11 (7840920)	0.0001
7	0.02 (2520)	0.4745	137.12 (3176460)	0.2247		
8	0.23 (20160)	0.4917	11662.00 (203222880)	0.2569		
9	2.51 (181440)	0.5018				
10	29.78 (1814400)	0.5113				
11	378.45 (19958400)	0.5176				
12	5265.00 (239500800)	0.5236				

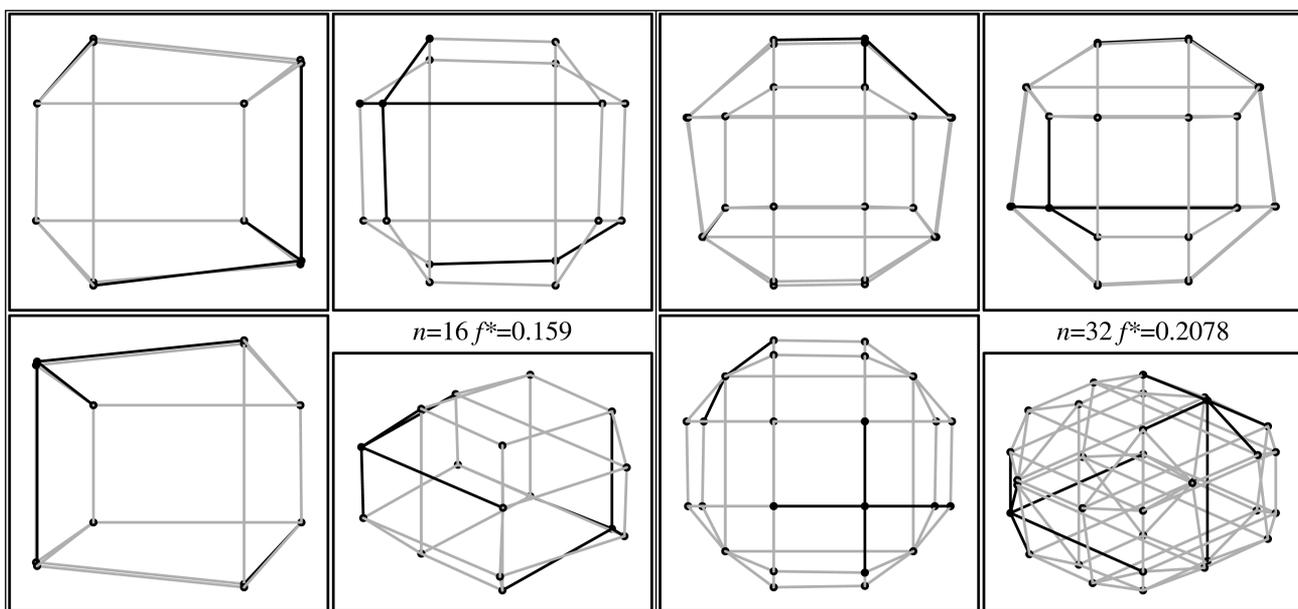


Fig 1. Projections of three-dimensional images of multidimensional cubes

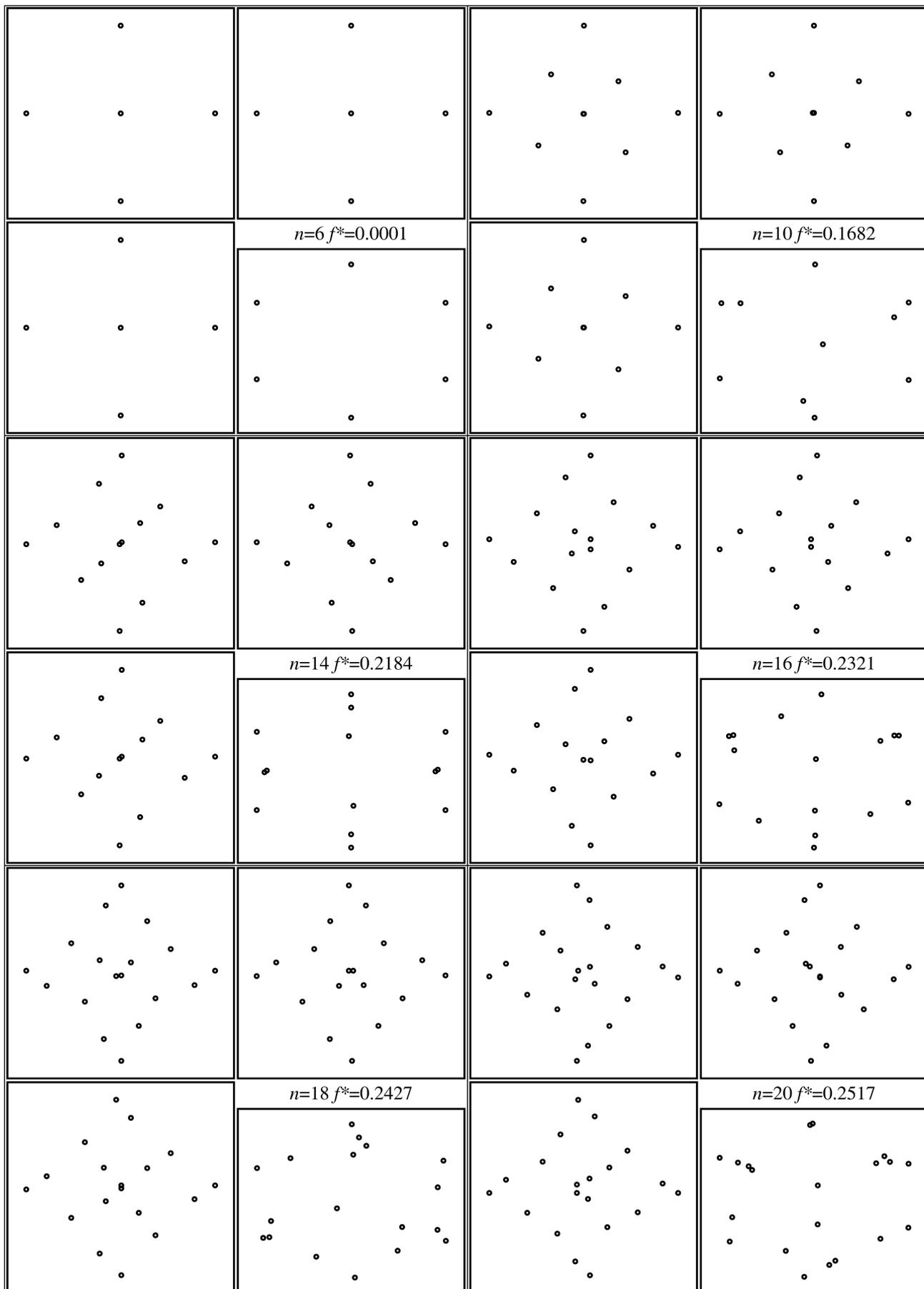


Fig 2. Projections of three-dimensional images of multidimensional equidistant simplexes

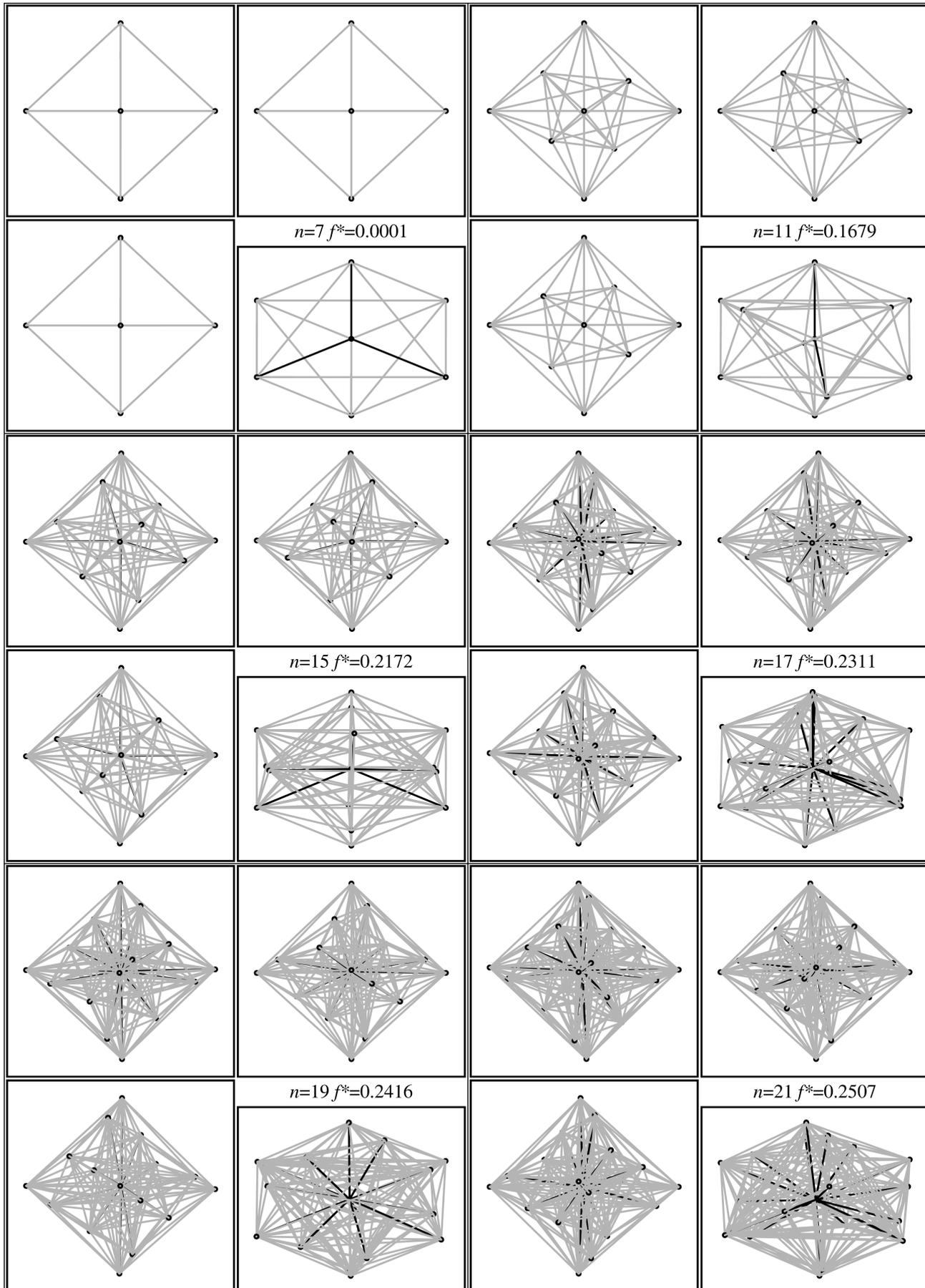


Fig 3. Projections of three-dimensional images of multidimensional right-angle simplexes

Lines adjacent to two opposite vertices are darker. The number of vertices (n) and values of relative errors (f^*) are given. The orthogonal projection on the yz -plane of three-dimensional image of four-dimensional cube is very similar to two-dimensional image of it given in [7], but other projections show the volume of three-dimensional image. Vertices form quadrangles representing sides of multidimensional cubes.

Projections of three-dimensional images of vertices of multidimensional equidistant simplexes are shown in Fig 2. The vertices are visualized as circles. No joins are shown as all vertices are adjacent to each other and all distances are equal in the original space. The vertices are visualized on the surface of diamond-shaped structure whose orthogonal projections are rhombs. The points on the surface of such structure are of the same distance to the center when the city block metrics is used, therefore all vertices are shown alike.

Projections of three-dimensional images of vertices of multidimensional right-angle simplexes are shown in Fig 3. The vertices are visualized as circles. Joins adjacent to the “zero” vertex are darker than joins between the other vertices. The orthogonal projections are a little bit similar to two-dimensional images of simplexes given in [7], however three-dimensional images have volume. The “zero” vertex is always visualized in the center of three-dimensional image. The other vertices are visualized similarly as vertices of multidimensional equidistant simplexes – on the surface of diamond-shaped structure whose projections are rhombs, therefore all “non-zero” vertices are shown alike.

4. Conclusions

Multidimensional scaling with city block distances is a prospective technique for three-dimensional visualization of multidimensional data, e.g. three-dimensional images of geometric multidimensional objects highlight the essential geometric properties of the originals. The hybrid method, based on a combination of genetic search with quadratic

programming, is a practically acceptable heuristic method for solution of difficult optimization problem occurring in implementation of the considered version of multidimensional scaling.

Relative error of visualization increases when dimensionality of original data increases. Relative error of visualization decreases when dimensionality of embedding space increases. Time, required to solve quadratic programming problems at lower level of visualization algorithm, varies depending on original data.

5. Acknowledgements

The research is supported by Lithuanian State Science and Studies Foundation and the NATO Reintegration grant CBP.EAP.RIG.981300.

References

1. Borg, I.; Groenen, P. *Modern Multidimensional Scaling*. NY: Springer, 1997.
2. Törn, A.; Žilinskas, A. *Global Optimization. Lecture Notes in Computer Science*, Vol 350, 1989, p. 1–255.
3. Žilinskas, A.; Žilinskas, J. Two level minimization in multidimensional scaling. *Journal of Global Optimization*, 2006, accepted.
4. Groenen, P. J. F.; Mathar, R.; 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.
5. 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.
6. Mathar, R.; Žilinskas, A. On global optimization in two-dimensional scaling. *Acta Applicandae Mathematicae*, Vol 33, 1993, p. 109–118.
7. Žilinskas, A.; Žilinskas, J. On multidimensional scaling with Euclidean and city block metrics. *Technological and Economic Development of Economy*, Vol 12, No 1, 2006, p. 69–75.

DAUGIAMETIŲ DUOMENŲ VIZUALIZAVIMAS NAUDOJANT TRIMATES SKALES

A. Žilinskas, J. Žilinskas

Santrauka

Daugiamatės skalės naudojamos artimumu apibrėžtiems duomenims atvaizduoti taškais mažo mato erdvėje. Uždavinys sprendžiamas optimizuojant atstumo tarp atitinkamų taškų atitiktus duotiems artimumams įvertį. Vaizdavimo tikslumo priklausomybė nuo skalės ir duomenų mato yra aptarta ir pasiūlyta naudoti trimatės skalės. Įvairių duomenų trimatės skalės pavaizduotos projekcijomis ir aptartos.

Reikšminiai žodžiai: daugiamatės skalės, globali optimizacija, metaeuristika, miesto kvartalo metrika, daugiamatė duomenų vizualizavimas.

Antanas ŽILINSKAS. Professor and head of Department of Applied Informatics of VMU and principal researcher of the 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 fellow at University College of London (UK, 12 months). Research interests: global optimization, data analysis and parallel computing.