Communication & Cognition
Vol. 43, Nr. 1 & 2 (2010), pp. 3 - 13

A CONJUGATE GRADIENT METHOD FOR TWO DIMENSIONAL SCALING

Antanas Zilinskas, Audrone Jakaitiene

Abstract

Multidimensional scaling is a technique for representing multidimensional data (a set of points in a multidimensional space) in a space of lesser dimensionality. The case of the 2–dimensional embedding space is of a special interest since 2-dimensional images are well suitable for visualization. The quality of visualization is measured by the difference between the pair wise distances in the original and embedding spaces defined by the *STRESS* function. The latter should be minimized. This complicated (multimodal) minimization problem can be tackled by a hybrid method combining a genetic type algorithm with a conjugate gradient descent routine. In the present paper a version of conjugate gradient method oriented to *STRESS* minimization is considered.

Keywords

Visualization, optimization, multidimensional scaling, local descent, conjugate gradient method, convergence

1. Introduction

Visualization is a technique for the heuristic analysis of multidimensional data where multidimensional points are mapped to 2–dimensional plane preserving the structure of pair wise distances. Such a mapping is a special case of multidimensional scaling (MDS). The problem of MDS was formulated in (Kruskal, 1964, pp. 1-27) as a problem of minimization of *STRESS* function. The term "*multidimensional scaling"* was accepted by statisticians and users of this technique in social sciences (Mathar, 1995, p. 44), (Borg, Groenen, 1997). The paper by Sammon (Sammon, 1969, pp. 401-409) originated the development of a special version of MDS. The subsequent papers on implementation and application of Sammon's method use the term "*nonlinear mapping technique"*; see e.g. (Dzwinel, 1994, pp. 949-

959). Applications of Sammon's method are normally oriented to the problems of pattern recognition. Although different aspects of MDS have been investigated, crucial implementation difficulties remain not resolved.

2. Formulation of MDS problem

Let us give a short formulation of the problem. The matrix δ_{ij} , $i, j = 1,...,n$ gives the pair wise dissimilarities between *n* objects, and it is symmetric. Dissimilarities are data for MDS; for example, they can be obtained experimentally. In case the original data is a set of points in multidimensional space R^d the dissimilarities are defined as distances in this space. The points $x_i \in R^m$, $i = 1,...,n,$ should be found which inter-point Euclidean distances fit the given dissimilarities. The embedding Euclidean space normally is 2-dimensional (*m*=2), but other dimensionalities may be also interesting for some applications, (*m<d).* To find the points x_i the *STRESS* function *f (X)* should be minimized,

$$f(X) = \sum_{i<j} w_{ij} (d_{ij}(X) - \delta_{ij})^2, \quad X = (x_{11},..., x_{n1}, x_{12},..., x_{nm}) \tag{1}$$

where $d_{ij}(X)$ denotes the Euclidean distance between the points x_i , x_j. It is supposed that the weights are positive: $w_{ij} > 0,\ i, j = 1,...,n.$

Although the *STRESS* function is defined by the analytical formula (1), which seems rather simple, it normally has many local minima. The minimization problem is highly dimensional, with the number of variables equal to *N*=*n*×*m*. At some points the function f*(X)* is not differentiable. The listed features make minimization of f*(X)* difficult.

3. Local minimization

It is well known (De Leeuw, 1984, pp. 111-113) that *f(X)* is differentiable at a local minimum point, i.e. if *X* is a local minimizer, then the following equalities and inequalities are valid

$$\frac{\partial f(X)}{\partial x_{kh}} = 2\sum_{i \neq k} w_{ki}(1 - \frac{\delta_{ki}}{d_{ki}(X)})(x_{kh} - x_{ih}) = 0, d_{ij}(X) > 0,\ i, j = 1,..., n.$$

In fact, the more general result may be proven:

Proposition. Let *L(t), -∞<t<∞,* be a line in R^N , containing a point at which *f(•)* is differentiable. Then *f(•)* is differentiable at any point *L(t*)*$\in R^N$, where *t** is the local minimizer of *φ(t)=f(L(t)).*

It follows from the proposition that a local descent trajectory escapes the points of non–differentiability of *f(•)*. Therefore, a fast local descent method may be applied to find a local minimizer of *f(•).*It is well known that variable metric methods are efficient for local minimization of f*(X)* in case of not too high dimensionality *N.* For very high dimensionalities a conjugate gradient method seems promising. From a theoretical point of view the rate of convergence is most important feature of an algorithm. The super linear convergence rate of a conjugate gradient method may be proved under mild assumptions on an objective function. However, the quadratic convergence is proved assuming the norm of Hessian of the objective function be bounded from zero. Let us analyze the second directional derivative of *f(X)* at a local minimum point *Z* with respect to the direction *S*:

$$\varphi(t,S) = f(Z + t \cdot S),\ \varphi'(0,S) = 0,$$

$$\varphi'(0,S) = 2\sum_{i<j} w_{ij}\left(1 - \frac{\delta_{ij}}{d_{ij}(Z)}\right)\sum_{h=1}^{m}(z_{ih} - z_{jh})(s_{ih} - s_{jh}),$$

$$\varphi''(0,S) = 2\sum_{i<j} w_{ij}\{d_{ij}^2(S) - \frac{\delta_{ij}}{d_{ij}^3(Z)}(d_{ij}^2(S) \cdot d_{ij}^2(Z) - [\ \sum_{h=1}^{m}(z_{ih} - z_{jh})(s_{ih} - s_{jh})\ ]^2)\}.$$

At a local minimum point the inequality $\varphi''(0,S) \geq 0$ holds for any *S*. For the directions *S* corresponding to the translations and rotations of the embedding space the inequality is reduced to the equality$\varphi''(0,S) = 0$. The latter equality implies the degeneracy of Hessian. To ensure the quadratic convergence of the conjugate gradient method the problem should be regularized. A simple regularization via the fixation of several variables to exclude invariance with respect to translations and rotations of the embedding space has some serious disadvantages; see e.g. (Žilinskas, 1997, pp. 200-204).. The regularization can be achieved also by excluding invariance with respect to translations and rotations (in the 2–dimensional case, *m=2*) introducing the equality constraints:

$$\sum_{i=1}^{n} x_{i1} = \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} x_{i1} x_{i2} = 0. \qquad (2)$$

Proposition The Polak – Ribiere conjugate gradient method converges to a local minimizer of regularized minimization problem quadratically with respect to the number of iterations including *2n-3* exact line searches.

Since the analytic expressions of the first and second directional derivatives are available, then a high precision line search method based on forth degree polynomial interpolation may be easily implemented.

4. A hybrid method

Several algorithms of minimization of the *STRESS* are available. The theoretical results of the previous section show that the non–differentiability of (1) is not a concern if a local descent method is used. Therefore, a conjugate gradient method with regularization of minimization problem is a strong competitor for other well known methods. The latter method may successfully cope with high dimensionality. Therefore, the real difficulty is caused by the multimodality of the *STRESS.*

The majorization method, which is especially tailored for MDS, may escape some local minima. However, like the other local methods, it provides a solution essentially depending on a starting point (Borg, Groenen, 1997). A local descent may be extended for multimodal problems using the tunneling approach. The possibilities of such an extension for MDS are discussed in (Borg, Groenen, 1997). The combination of majorization method with genetic algorithm was proposed in (Mathar, 1996, pp. 63-71) where the method of Mathar and Žilinskas (Mathar, Žilinskas, 1993, pp. 109-118) was modified substituting a variable metric local descent with majorization method. The results of limited experimental testing showed that a genetic type approach may be promising for global minimization of (1). In the references on “nonlinear mapping” various versions of descent are of primary interest. However, combinations of descent with general global search methods, e.g. simulated annealing, are claimed promising in (Dzwinel, 1994, pp. 949-959).

The proposed version of the conjugate gradient algorithm is supposed to be combined with a genetic algorithm. A brief description of this genetic algorithm is presented below; for more details we refer to (Zilinskas and Zilinskas, 2008, pp.429-443). The majority of the authors consider the unconstrained minimization of (1). In such a case the same minimum value may be obtained at different points on the orbit of local minimizers corresponding to invariance of f*(X)* with respect to translation and rotation of the embedding space. Therefore several copies of the same local minimum may be obtained, but with very different minimizers. It is difficult to handle such information on local minima to perform rational search for a global minimizer. In our conjugate gradient algorithm the ortogonalization (2) prevents multiplication of the minimizers. The analysis carried out for several sets of data shows that different local minimizers are located rather close each to other. Geometrically this feature of the *STRESS* may be explained as a small structural difference between the graphs in the 2–dimensional embedding space corresponding to different local minimizers of *f(X).* The geometric interpretation implies a hypothesis that graphs corresponding to local minimizers are composed of similar semi-optimal sub-graphs.

In terms of the genetic algorithms different local minima are considered as the ideal representatives of different breeds. They are used for crossover. The initial population of the size *p* is generated by means of local descent method from the random initial points. The parents are chosen at random with uniform distribution. The graph in the embedding space is considered the phenotype. The hypothesis is accepted that crossover of chromosomes imply the crossover of characteristics of phenotype. The latter is modeled as the crossover of the graphs in the embedding space, i.e. some points $x_i \in R^2$ of one parent graph and remaining points of the other parent graph are taken to compose the descendant graph. The break position is generated randomly with uniform distribution in the interval [1, *n*]. The larger number of points is taken from the fittest parent. The mutation is modeled as the random summands to the components of graph coordinates; the distribution of the random variable is uniform in the interval [-*r*, *r*]. Two selection mechanisms have been investigated: 1) the descendants survive and the parents die; 2) each descendant competes with a randomly chosen individual of current population. The number of generations modeled to find the global minimum is denoted *g.*

5. Experimental testing

To assess the efficiency of the regularized conjugate gradient method (CG) for local minimization of (1) CG was compared with the known method based on majorization approach (MA) (Mathar, pp. 63-71). The convergence of MA to local minimum point is proved, e.g. in (Mathar, 1995). Moreover, it may be expected that bad local minima will be avoided. Both methods (CG and MA) were implemented in MATLAB. Two known sets of data were used. The first set, presented in Table 1 a), contains δ_{ij} for ten soft drinks whose dissimilarities are obtained by means of experimental testing (Borg, Groenen, 1997). The second set of data corresponding to the proximities of 13 facial expressions (Borg, Groenen, 1997) is presented in Table 1 b).

Table 1. Data for visualization.

a)

1.27	1.69	2.04	3.09	3.20	2.86	3.17	3.21	2.38
	1.43	2.35	3.18	3.22	2.56	3.18	3.18	2.31
		2.43	3.26	3.27	2.58	3.18	3.18	2.42
			2.85	2.88	2.59	3.12	3.17	1.94
				1.55	3.12	1.31	1.70	2.85
					3.06	1.64	1.36	2.81
						3.00	2.95	2.56
							1.32	2.91
								2.97

b)

0.405											
0.825	0.254										
0.557	0.269	0.211									
0.115	0.267	0.898	0.378								
0.297	0.388	0.927	0.605	0.234							
0.434	0.853	1.187	0.978	0.712	0.136						
0.490	0.131	0.256	0.421	0.590	0.518	0.847					
0.625	0.188	0.074	0.045	0.477	0.545	1.020	0.263				
0.155	0.484	0.925	0.492	0.222	0.417	0.544	0.545	0.710			
0.168	0.581	0.792	0.542	0.434	0.472	0.431	0.379	0.658	0.198		
0.657	0.743	0.830	0.893	0.816	0.466	0.157	0.649	0.977	0.493	0.483	
0.393	0.451	0.847	0.348	0.160	0.489	0.918	0.605	0.655	0.412	0.351	1.265

To compare the performance of both methods 100 runs have been performed with both sets of test data. The starting points were generated randomly with uniform distribution in the cube $[-1.2, 1.2]^N$. Both methods used the same random starting point and the same stopping condition: $\| \nabla f(X_k) \| = \varepsilon < 0.001$. The average values of the results are presented in Table 2.

Table 2. The comparison of performance of CG and MA algorithms

	STRESS	Time (s)	Iterations
CG, Test 1	13.18	2	36
MA, Test 1	13.30	5	213
CG, Test 2	0.684	2	22
MA, Test 2	0.684	3	77

The known best value of the *STRESS* with the first set of data is 11.746, and region of attraction of the corresponding minimizer makes 4% of the feasible region. By means of CG a minimizer with the value better than 11.75 has been found five times, and by means of MA has been found once. 75 times the value found by means of CG was better than the value found by means of MA.

The second set of data defines the *STRESS* function, which is very likely unimodal. All 100 runs for both methods stopped in the vicinity of the same local minimizer. Average time of minimization of this rather simple function by CG is again considerably better than by MA. In all runs a found function value was slightly better for CG than for MA.

Time of local descent to a local minimizer depends on dimensionality of a problem and on the stopping condition. The dimensionality of minimization problems corresponding to Test 1 and Test 2 is equal to 20 and 26 correspondingly. The third test problem is 2-dimensional representation of vertices of the 5-dimensional cube. The vertices are numbered according their digital representation. Since the number of vertices is equal to 32, then the dimensionality of the minimization problem is equal to 64. The local descent for the problems of such a dimensionality is time consuming. In the Table 3 the results of two runs are presented. The results show that the time of local descent for 64-dimensional problem is considerably larger than the time for the 20-dimensional cases. The solution with the prescribed tolerance of gradient norm by means of MA was not found.

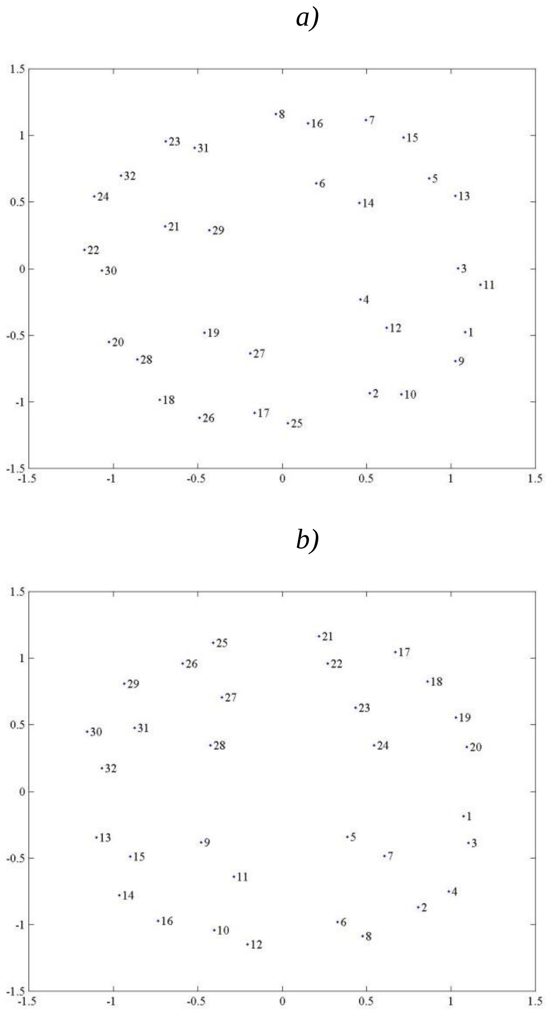
Table 3. Dependence on the stopping condition

	Time (s)	Iterations	Time (s)	Iterations
	$\varepsilon = 0.001$		$\varepsilon = 0.01$	
CG	35	73	32	63
MA	228	1000	228	1000

Since the CG has been shown faster than MA, the former was used to construct a global search algorithm. The size of population, number of generations, crossover method, and mutations intensity has to be chosen experimentally. The experiments were performed with the Test 1. Summarizing the results, the following parameters may be recommended: p=10, g=3, competition selection mechanism, r=0.3. The averaged results of 100 runs for Test 1 are the following: the *STRESS* value equal to 11.77, number of line search iterations equal to 812, minimization time equal to 35. Minimal function value with accuracy no less than 1% was found 95 times. The average results obtained by means of evolutionary algorithm are considerably better than those obtained by means of local descent from the random initial points.

The evolutionary algorithm was applied also to Test 3. The best-found value of *STRESS* function was 141.11, number of line search iterations was 1470, and solution time was 704. The 2–dimensional image of vertices, corresponding to the global minimizer, is presented in Fig.1 a). In Fig.1 b) the image corresponding to a local minimizer is presented. Although the difference of the *STRESS* values is insignificant, there are clearly visible differences of the images.

The 2–dimensional images of the results on search process may be useful for understanding of character of multimodal objective functions. In Fig.2 the images of trajectories of local search from seven random initial points (squares) are presented. The function (1) with the data of Test 1 was minimized. The picture supports the hypothesis that different local minimizers of (1) are close each to other.



a)

b)

Fig 1. 2-dimensional image of vertices of 5–dimensional cube:
a) global minimum - 141.11; b) local minimum - 141.63.

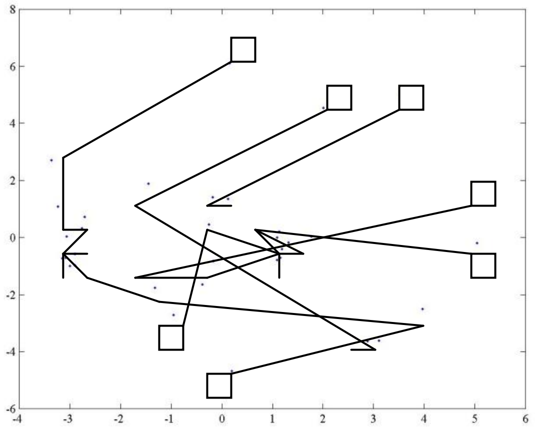


Fig 2. Two-dimensional image of descent trajectories in 20 - dimensional space

6. Conclusions

Genetic type global optimization algorithms are prospective to solve a difficult global minimization problem of MDS. The latter is useful for visualization of information on a global search process.

Acknowledgements

This research was partially supported by Agency for International Science and Technology Development Programmes in Lithuania through COST IC0602 programme.

References

Borg I., and P. Groenen, 1997. *Modern Multidimensional Scaling,* NY: Springer.

De Leeuw J. 1984. "Differentiability of Kruskal's *STRESS* at a local minimum", *Psychometrika,* Vol.49, 111-113.

Dzwinel W. 1994. "How to make Sammon's mapping useful for multidimensional data structures analysis", *Pattern Recognition* Vol.27, No.7, 949-959.

Kruskal J. 1964. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". *Psychometrika,* Vol.29, 1-27.

Mathar R. 1996. A hybrid global optimization algorithm for multidimensional scaling, *Classification and Knowledge Organization,* Eds. R.Klar and O.Opitz, Springer, 63-71.

Mathar R. 1995. *Multidimensionale Skalierung, Mathematische Grundlagen und Algorithmische Konzepte,* Preprint, RWTH Aachen, 44p.

Mathar R., and A. Žilinskas. 1993. "On global optimization in two – dimensional scaling", *Acta Applicandae Mathematicae,* Vol.33, 109-118.

Sammon J. 1969. "A nonlinear mapping for data structure analysis", *IEEE Trans. Comput.* Vol.C-18, 401-409.

Törn A. and A. Žilinskas 1989. *Global Optimization,* Berlin: Springer.

Žilinskas A. 1997. "On quadratic convergence of a visualisation algorithm". *Proc. of 4th International Conference on Pattern Recognition and Information Processing, Minsk-Szczecin,* Vol.1, 200-204.

Žilinskas A. and Žilinskas J., (2008) A hybrid method for multidimrnsional scaling using city block distances, *Math. Meth. Oper. Res.,* Vol.68, 429-443.

About the authors

Principal researcher Antanas Zilinskas, Prof. Dr. Habil., Optimization Sector, Institute of Mathematics and Informatics, Phone: +370 5 21 09 338, E-mail: antanasz@ktl.mii.lt.

Senior researcher Audrone Jakaitiene, Dr., Optimization Sector, Institute of Mathematics and Informatics, Phone: +370 5 21 09 304, E-mail: audrone.jakaitiene@ktl.mii.lt.