INFORMATICA, 1996, Vol. 7, No. 2, 268–274

A QUADRATICALLY CONVERGING ALGORITHM OF MULTIDIMENSIONAL SCALING

Antanas ŽILINSKAS

Institute of Mathematics and Informatics
2600 Vilnius, Akademijos St. 4,
Vytautas Magnus University
3600 Kaunas, Vileikos St. 8, Litnuania

Abstract. Multidimensional scaling (MDS) is well known technique for analysis of multidimensional data. The most important part of implementation of MDS is minimization of *STRESS* function. The convergence rate of known local minimization algorithms of *STRESS* function is no better than superlinear. The regularization of the minimization problem is proposed which enables the minimization of *STRESS* by means of the conjugate gradient algoritm with quadratic rate of convergence.

Key words: local minimization, conjugate gradients, quadratic convergence rate.

1. Introduction. Multidimensional scaling (MDS) is a technique for analysis of multidimensional data widely usable in different applications (Mathar, 1995; Wish and Carrol, 1982). The theoretical and algorithmic aspects of MDS are considered, e.g., in (Groenen, 1993; Mathar, 1989; de Leeuw and Heiser, 1982; Mathar, 1995). Let us give a short formulation of the problem. The pairwise dissimilarities between n objects are given by the matrix $(\delta)_{ij}$, $i,j = 1,\ldots,n$. The points $x_i \in R^n$, $i = 1,\ldots,n$ should be found which interpoint Euclidean distances fit given dissimilarities. The embedding Euclidean space R^m normally is two dimensional ($m = 2$), but the other dimensionalities may be also interesting for some applications. To find the points x_i the *STRESS* function $f(x)$ should be minimized, where

$$f(x) = \sum_{i<j}^{n} w_{ij}\left(d_{ij}(X) - \delta_{ij}\right)^2, \tag{1}$$

$X = (x_{11},\ldots,x_{n1},x_{12},\ldots,x_{nm})$ and $d_{ij}(X)$ denotes the Euclidean distance between the points x_i and x_j. It is supposed that the weights are positive:

$w_{ij} > 0,\ i,j = 1 \ldots n$. On one hand, MDS is related to global optimization as a difficult problem (Mathar and Žilinskas, 1993, 1994) and on the other hand as a tool for visualization of the optimization process (Žilinskas, 1993). Even the local minimization of $f(x)$ is not easy: normally the problems are of large dimensionality $N = n \times m$, the objective function is not everywhere differentiable. The known methods have the rate of convergence no better than superlinear (Mathar, 1989, 1995). In the present paper some new features of $f(x)$ are proved and regularization of the minimization problem is proposed enabling the local minimization of $f(x)$ with quadratic rate of convergence.

2. Local descent trajectory escapes of nondifferentiability. The following statement generalizes the well known result (de Leeuw, 1984) that $f(x)$ is differentiable at a local minimum point.

Theorem 1. *Let $L(t),\ -\infty < t < \infty$ be a line in R^N, containing point E at which $f(\cdot)$ is differentiable. Then $f(\cdot)$ is differentiable at any point $L(t^*) \in R^N$ where t^* is local minimum point of $\phi(t) = f(L(t))$.*

Proof. The function $f(\cdot)$ is differentiable everywhere except the points satisfying the condition $d_{kl}(\widetilde{X}) = 0$, i.e., $\widetilde{x}_{kr} = \widetilde{x}_{lr},\ r = 1, \ldots, m$. If the point $\widetilde{X}$ is on the line $L(t)$, then $L(t)$ may be represented as

$$L(t) = \{X\colon\ x_{ir} = \widetilde{x}_{ir} + t \cdot (e_{ir} - \widetilde{x}_{ir}),\\ i = 1, \ldots, n,\ r = 1, \ldots, m\},$$

and the function $\phi(t) = f(L(t))$ may be expressed as follows

$$\phi(t) = \sum_{(i,j) \in I_{kl}} w_{ij}(d_{ij}(L(t)) - \delta_{ij})^2 + w_{kl}(d_{kl}^2(L(t)) + \delta_{kl}^2 \\ - 2w_{kl}\delta_{kl}d_{kl}(L(t))\Psi_{kl}(t) - a_{kl}|t|,$$

where I_{kl} is set of indexes $\{(i,j),\ i < j\}$ without k, l, $\Psi_{kl}(t)$ is continously differentiable at the point $t = 0$ and $a_{kl} > 0$. A point of local minimum of function $\Psi_{kl}(t) - a_{kl}|t|$, never coincides with $t = 0$: if $|\Psi'_{kl}(t)| > a_{kl}$ then $\phi(t)$ is increasing or decreasing at $t = 0$, if $|\Psi'_{kl}(t)| \leqslant a_{kl}$ then $\phi(t)$ attains a local maximum at $t = 0$. Therefore, the point $\widetilde{X} = L(0)$ at which function $f(X)$ is not differentiable never coincides with a point of local minimum of $f(X)$ on a line.

COLLORARY. Let a local descent method defined by the recurrent formula $X^{k+1} = X^k + t_k S^k$ where $f(X^{k+1} = \min_t f(X^k + tS^k)$ is started at the point of differentiability of $f(X)$. Then no point of the sequence X^k coincides with any point of non differentiability of $f(X)$.

3. Minimization of *STRESS* by an algorithm of conjugate gradients. Since a trajectory of local descent started at the point of differentiability of $f(x)$ escapes of the points of nondifferentiability, then an algorithm with a high convergence rate may be applied to minimize $f(x)$. The quasi Newton methods normally are not considered as the prospective candidates because of large dimensionality of the practical problem. For the test problems of modest dimensionality quasi Newton methods perform very well (Mathar and Žilinskas, 1993). With respect to the convergence rate, quasi Newton and conjugate gradient methods are similar, but the later are more suited to solve problems of large dimensionality. The standard version of the conjugate gradients algorithm (Poliak, 1988) is as follows:

$$X^{k+1} = X^k + t_k S^k, \quad t_k = \operatorname*{argmin}_{t>0} f(X^k + tS^k), \tag{2}$$

$$S^k = G^k + \beta_k S^{k-1}, \quad \beta_k = \frac{(G^k, G^k - G^{k-1})}{||G^{k-1}||^2},$$
$$G^k = -\nabla f(X^k), \quad \beta_0 = 0.$$

For the one dimensional minimization of $\varphi(t) = f(X^k + tS^k)$ the Newton's method may be applied

$$t_{\nu+1} = t_\nu - \frac{\varphi'(t_\nu)}{\varphi''(t_\nu)}. \tag{3}$$

However, we do not have a proof that $\varphi''(t) > 0$ along all descent trajectories. Therefore, if during the one-dimensional minimization the condition $\varphi''(t) \leqslant 0$ would be satisfied, then the Newton's method would be changed to the more sophisticated one dimensional algorithm. To simplify the formulas of derivatives in (3) let us denote the point in R^N corresponding to t_ν by X and the descent direction by S. Then only the derivative values at point $t = 0$ are used in (3):

$$\varphi'(0) = 2\sum_{i<j} Big(1 - \frac{\delta_{ij}}{d_{ij}(X)}\Big) \sum_{h=1}^{m}(x_{ih} - x_{jh})(s_{ih} - s_{jh}), \quad (4)$$

$$\varphi''(0) = 2\sum_{i<j}\sum_{h=1}^{m}(s_{ih} - s_{jh})^2 - \frac{\delta_{ij}}{d_{ij}(X)} \frac{(x_{ih} - x_{jh})^2(s_{ih} - s_{jh})^2}{d_{ij}^2(X)}. \quad (5)$$

4. Rate of the convergence. It is well known (Poliak, 1988) that the method (2) converges to a local minimum with the quadratic rate if the following conditions are satisfied:

1) the Hessian of $f(X)$ is positively defined at the local minimum point,
2) the Hessian of $f(X)$ satisfies the Lipshits condition in the vicinity of local minimum point.

The first of these conditions can not be satisfied for the function $f(X)$ because it is invariant with respect of translation: $f(X) = f(X + C)$ for every $C = (c_1, \ldots, c_1, \ldots, c_m, \ldots, c_m)$. The equality $\varphi(t) = f(X + tC) = \text{const}$ implies the equality $\varphi''(t) = -c\nabla^2 f(X)C = 0$ and $C\nabla^2 f(X)C = 0$, which hold also for a local minimum point. Let us regularise the minimization problem supposing $x_{11} = x_{12} = \ldots = x_{1m} = 0$. Then the dimensionality of minimization problem is decreased by m and the invariance in respect of translation is excluded. We will show that the Hessian becomes positively defined at the local minimum point X, satisfying the condition of non degeneracy of embedding: for each h there exist the pair (i, j) such that $x_{ih} \neq x_{jh}$. This condition means that the embedding, defined by the local minimum point X, can not be arranged in the subspace of lower dimentionality than m.

Let X be a local minimum point of $f(\cdot)$ in respect of X with $x_{11} = x_{12} = \ldots = x_{1m} = 0$, and let S be a direction in R^N satisfying the condition $s_{11} = s_{12} = \ldots = s_{1m} = 0$. Since X is a local minimum point of $f(\cdot)$ then 0 is local minimum point of $\varphi(t) = f(X + tS)$ implying:

$$\varphi'(0) = 2\sum_{i<j}\Big(1 - \frac{\delta_{ij}}{d_{ij}(X)}\Big)\sum_{h=1}^{m}(x_{ih} - x_{jh})(s_{ih} - s_{jh}) = 0. \quad (6)$$

On the other hand

$$\varphi''(0) = 2\sum_{i<j}\Big(1 - \delta_{ij}d_{ij}(X)\Big)\sum_{h=1}^{m}(s_{ih} - s_{jh})^2 +$$

$$+2\sum_{i<j}\frac{\delta_{ij}}{d_{ij}^3(X)}\sum_{h=1}^{m}(s_{ih}-s_{jh})^2(x_{ih}-x_{jh})^2. \qquad (7)$$

Since (6) is valid for any direction S, it is valid also for $S = X$. Therefore the first summand in (7) is equal to 0. The second summand is non negative implying $\varphi''(0) \geqslant 0$. Let us suppose $\varphi''(0) = 0$, then

$$(s_{1h}-s_{jh})^2(x_{1h}-x_{jh})^2 = 0, \quad j = 2,\ldots,n,\ h = 1,\ldots,m. \qquad (8)$$

Since $d_{1j} \neq 0,\ j = 2,\ldots,n$ then for each j there exists h such that $s_{jh} = s_{1h} = 0$. Let us include into the list I_h the indexes j such that $s_{jh} = 0$. It was supposed that for each h there exist such a pair (i,j) that $x_{ih} \neq x_{jh}$, implying the existence of $x_{i(h),h} \neq 0, h = 1,\ldots,m$, and $i(h) \in I_h$. Therefore, each list I_h contains at least two elements: $1, i(h) \in I_h,\ h = 1,\ldots,m$. Let us suppose $i(h) = 2$, i.e. $x_{2h} \neq x_{1h}$. Then for $i \notin I_h$ there holds the inequality $x_{ih} \neq x_{2h}$. However, the equality $(s_{2h}-s_{ih})^2(x_{2h}-x_{ih})^2 = 0$ should be satisfied for all $i = 3,\ldots,n$, and $x_{ih} \neq x_{2h}$ implies $s_{2h} = s_{ih} = 0$, and finally $s_{ih} = 0,\ i = 1,\ldots,n$.

If $i(h) = n$ then $(s_{nh}-s_{ih})^2(x_{nh}-x_{ih})^2 = 0,\ i = 2,\ldots,n-1$ implies $s_{ih} = 0,\ i \notin I_h$, and finally $s_{ih} = 0,\ i = 1,\ldots,n$. A bit more long but similar analysis of the case $2 < i(h) < n$ completes the proof that the assumption $\varphi''(0) = 0$ implies the equality $||S|| = 0$. Since $\varphi''(0) = S\nabla^2(X) > 0$, $||S|| > 0$, then the Hessian of regularised *STRESS* function is positively defined.

To show that the Hessian of *STRESS* satisfies Lipshitz condition in the vicinity of local minimum point it is sufficient to show that the second derivatives of the function satisfy Lipshitz condition under the same assumption. At the local minimum point X the inequality $d_{ij} > 0$ holds implying that all the following functions

$$\frac{\partial^2 f(X)}{\partial x_{kh}^2} = 2(n-1)+2\sum_{j\neq k}\frac{\delta_{kj}((x_{kh}-x_{jh})^2-d_{kj}^2(X))}{d_{kj}^3(X)},$$

$$\frac{\partial^2 f(X)}{\partial x_{kh}x_{sh}} = -2+2\frac{\delta_{ks}(-(x_{kh}-x_{sh})^2+d_{ks}^2(X))}{d_{ks}^3(X)},$$

$$\frac{\partial^2 f(X)}{\partial x_{kh}x_{sr}} = -2\frac{\delta_{ks}(x_{kh}-x_{sh})(x_{kr}-x_{sr})}{d_{ks}^3(X)}$$

are Lipshitzian. Summarazing the conclusions, the following theorem may be formulated.

Theorem 2. *The conjugate gradient algorithm* (2) *converges to a non degenerated local minimum of regularised STRESS with the quadratic rate of convergence.*

REFERENCES

Groenen, P.J.F. (1993). *The Majorization Approach to Multidimentional Scaling.* DSWO Press, Leiden.

Mathar, R. (1989). Algorithms in multidimensional scaling. In O. Opitz (Ed.), *Conceptual and Numerical Analysis of Data.* Springer, Berlin. pp. 159–177.

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

Mathar, R., and A. Žilinskas (1994). A class of test functions for global optimization. *Journal of Global Optimization,* **5**(2), 195–200.

Mathar, R. (1995). Multidimensionale skalierung, mathematische grundlagen und algorithmische konzepte. *Preprint,* RWTH Aachen.

de Leeuw, J., and W. Heiser (1982). Theory of multidimentional scaling. In P.R. Krishnaiah (Ed.), *Handbook of Statistics,* Vol. 2. North Holland, Amsterdam. pp. 285–316.

de Leeuw, J. (1984). Differentiability of Kruskd's stress at a local minimum. *Psychometrika,* **49**, 111–113.

Poliak, B. (1988). *Introduction into Optimization.* Nauka, Moscow (in Russian).

Wish, M., and J. Carroll (1982). Multidimensional scaling and its applications. In P.R.Krishnaiah (Ed.), *Handbook of Statistics,* Vol. 2. North Holland, Amsterdam. pp. 317–345.

Žilinskas, A. (1993). On visualization of optimization process. In J.Guddat et al. (Eds.), *Parametric Optimization and Related Topics 3,* Peter Lang Verlag, Frankfurt am Main. pp. 549–556.

Received April, 1996

A. Žilinskas received Candidate of Sciences (Ph.D.) degree from Kaunas University of Technology in 1973, and Dr. Habil. degree St.Petersburg University in 1985. He is a professor and a head of Informatics Chair at Institute of Mathematics and Informatics, and Vytautas Magnus University. He is a member of International Engineering Academy, AMS, IFIP, WG 7.6, IEEE CS, Fellow of American Bibliographical Institute. He is a member of editoral boards of "Journal of Global Optimization", "Control and Cybernetics", "Informatica". His research interest are optimization, visualization of multidimensional data and optimal design.

KVADRATINIO GREIČIO DAUGIAMAČIŲ SKALIŲ SUDARYMO ALGORITMAS

Antanas ŽILINSKAS

Vienu iš reikšmingiausių daugiamačių duomenų analizės metodų yra daugiamačių skalių metodas, kurį realizuojant reikia minimizuoti *STRESS* funkciją. Žinomų lokalios minimizacijos algoritmų *STRESS* funkcijai minimizuoti konvergavimo greitis yra ne geresnis negu supertiesinis. Šiame straipsnyje pasiūlyta reguliacija, leidžianti sujungtinių gradientų metodą pritaikyti *STRESS* minimizavimui, pasiekiant kvadratinį konvergavimo greitį.