An algorithm for realizing Euclidean distance matrices

Jorge Alencar [1]

Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas Gerais, Inconfidentes, MG, Brazil

Tibérius Bonates [2]

Depto de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Fortaleza, CE, Brazil

Carlile Lavor [3]

Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Campinas, SP, Brazil

Leo Liberti [4]

CNRS LIX, École Polytechnique, 91128 Palaiseau, France

Abstract

We present an efficient algorithm to find a realization of a (full) $n \times n$ squared Euclidean distance matrix in the smallest possible dimension. Most existing algorithms work in a given dimension: most of these can be transformed to an algorithm to find the minimum dimension, but gain a logarithmic factor of n in their worst-case running time. Our algorithm performs cubically in n (and linearly when the dimension is fixed, which happens in most applications).

Keywords: Distance geometry, sphere interesection, EDM, embedding dimension.

1 Introduction

The problem of adjustment of distances among points has been studied since the first decades of the 20th century [2]. It can be formally defined as follows: Let D be a $n \times n$ symmetric hollow (i.e., with zero diagonal) matrix with non-negative elements. We say that D is a squared Euclidean Distance Matrix (EDM) if there are $x_1, x_2, \ldots, x_n \in \mathbb{R}^K$, for a positive integer K, such that

$$D(i,j) = D_{ij} = \|x_i - x_j\|^2, i, j \in \{1, \ldots, n\},$$

where $\|\cdot\|$ denotes the Euclidean norm. The smallest K for which such a set of points exists is called the *embedding dimension* of D, denoted by $\mathsf{dim}(D)$. If D is not an EDM, we define $\mathsf{dim}(D) = \infty$.

We are concerned with the problem of determining $\mathsf{dim}(D)$ for a given symmetric hollow matrix D. If $\mathsf{dim}(D) = K < \infty$, we also want to determine a sequence $x = (x_1, \ldots, x_n)$ of n points in $\mathbb{R}^K$ such that D is the EDM of x. We emphasize that D is a full matrix.

In the literature we prevalently find efficient methods for solving a related problem, i.e. whenever K is given as part of the input (see e.g. [1,6,7]). Each of these algorithms can be used within a bisection search to determine the embedding dimension, incurring a multiplicative factor of $\mathcal{O}(\log(n))$ to their running time. These algorithms also require the embedding of a clique in $\mathbb{R}^K$, a task that demands $\mathcal{O}(K^3)$ time. If we assume K is not given as part of the input, then we can assume that K is $\mathcal{O}(n)$ (although tighter bounds exist [8]). Thus, any of these algorithms can be used within a bisection search for a total time of $\mathcal{O}(n^3 \log(n))$, in the worst case. We propose an algorithm which accomplishes the same task in $\mathcal{O}(n^3)$ time. If the embedding dimension is known, all algorithms (ours, as well as those in [1,6,7]) reduce to linear time in n.

Our algorithm is based on the problem of determining the intersection of K spheres in $\mathbb{R}^K$, where K varies during the algorithm. The problem of determining the intersections of spheres is well known, as are its applications (see, e.g. [2]). We numerically compare the algorithm with an existing technique

[1] Email: `jorge.alencar@ifsuldeminas.edu.br`
[2] Email: `tb@ufc.br`
[3] Email: `clavor@ime.unicamp.br`
[4] Email: `liberti@lix.polytechnique.fr`

available in the literature, and show that our approach is superior in terms of realization quality.

2 Some results about EDM

In this section we present the theoretical basis of our algorithm. Let $[n] = \{1, \ldots, n\}$ and $[n_1, n_2] = \{n_1, n_1+1, \cdots, n_2-1, n_2\}$. Furthermore, if $U, V \subseteq [n]$ such that $V = \{v_1, \ldots, v_{n_1}\}, U = \{u_1, \ldots, u_{n_2}\}$ and D is a $n \times n$ matrix, then $D(V, U) = (d_{ij})$ is the submatrix of D such that $d_{ij} = D(v_i, u_j)$ with $i \in [n_1]$ e $j \in [n_2]$. Given a positive integer n, we define $\{x_i\}_{i=1}^n = \{x_1, x_2, \ldots, x_{n-1}, x_n\}$.

Given a EDM of order n, the following lemma establishes that the embedding dimension of the given EDM is greater than the embedding dimension of any of its $(n-1)$-th principal submatrices by at most one.

Lemma 2.1 *Let D be a $(n+1) \times (n+1)$ EDM. If $\mathsf{dim}(D([n], [n])) = K$, then $\mathsf{dim}(D) \in \{K, K+1\}$.*

The next lemma ensures that, given a set S of n points that realizes the n-th principal submatrix of a EDM of order $n+1$ and embedding dimension at most m, S can be augmented into a realizing set for the full matrix.

Lemma 2.2 *Let D be a $(n+1) \times (n+1)$ EDM, and let $\mathsf{dim}(D) \leq m$. Additionally, let $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^m$ be a set of points which realizes $D([n], [n])$, the n-th principal submatrix of D. Then there exists $x_{n+1} \in \mathbb{R}^m$ such that $\{x_i\}_{i=1}^{n+1}$ realizes D.*

The following theorem establishes necessary and sufficient conditions for a $n \times n$ symmetric hollow matrix with non-negative elements to be a EDM. If this matrix is a EDM with $\mathsf{dim}(D) = K$, then there exists a set of points which realizes the given matrix such that K of them form a triangular structure in some sense, as explained below.

Theorem 2.3 *Let K be a positive integer and D be a $n \times n$ symmetric hollow matrix with non-negative elements. D is a EDM with $\mathsf{dim}(D) = K$ if and only if there exist $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^K$ and an index set $I = \{i_j\}_{j=1}^{K+1} \subseteq [n]$ such that*

$$\begin{cases} x_{i_1} = 0 \\ x_{i_j}(j-1) \neq 0, j \in [2, K+1] \\ x_{i_j}(i) = 0, j \in [2, K+1], i \in [j, K], \end{cases}$$

where $\{x_i\}_{i=1}^n$ realizes D.

The theorem can be proved by induction using the Lemmata 2.1 and 2.2. This induction process suggests an algorithm to verify whether or not a matrix D is a EDM and, if so, to determine an embedding in the least possible dimension. The procedure is shown in Alg. 1, and we refer to it as `edmsph`, from "EDM" and "sphere". The pseudocode makes use of a function $\mathsf{expand}(x)$ which endows point vectors in the sequence x with an additional zero component. We denote the sphere centered in $p \in \mathbb{R}^{K+1}$ with radius r by $\mathbb{S}^K(p, r)$.

Algorithm 1 $K = \texttt{edmsph}(D, x)$

$I = \{1, 2\}$
$K = 1$
$(x_1, x_2) = (0, \sqrt{D_{12}})$
for $i \in \{3, \ldots, n\}$ **do**
    $\Gamma = \bigcap_{j \in I} \mathbb{S}^K(x_j, D_{ij})$
    if $\Gamma = \varnothing$ **then**
        return ∞
    else if $\Gamma = \{p_i\}$ **then**
        $x_i = p_i$
    else if $\Gamma = \{p_i^+, p_i^-\}$ **then**
        $x_i = p_i^+$
        $x \leftarrow \mathsf{expand}(x)$
        $I \leftarrow I \cup \{i\}$
        $K \leftarrow K + 1$
    else
        error: $\mathsf{dim\ aff}(\mathsf{span}(x_I)) < K - 1$
    end if
end for
return K

We remark that, given K spheres in $\mathbb{R}^K$, we assume their centers are in general position, i.e. they span a $(K-1)$-dimensional affine space. Then we have at most two points in the intersection of these spheres: we have no point if the intersection is empty, one point if the intersection lies in the $(K-1)$-dimensional affine space generated by the centers, and two points if there are no points of the intersection in the $(K-1)$-dimensional affine space generated by the centers. Using trilateration on the appropriately indexed points guaranteed by Thm. 2.3, finding Γ in Alg. 1 requires solving a triangular linear system of order K, for $K \in [2, \texttt{dim}(D)]$, which can be carried out in time proportional to K^2 per iteration. This leads to a total time of $\mathcal{O}(n^3)$ in the worst case.

3 Numerical Experiments

In [1] Dattorro developed `isedm`, an algorithm for checking whether a given symmetric hollow matrix D with non-negative entries is a EDM. The algorithm was based on a relationship between PSMs and EDMs from [3].

In the first series of experiments, we used the instances proposed by Moré and Wu [5], consisting of structures with s^3 elements ($s \in \mathbb{N}$) positioned on the three-dimensional lattice defined by $\{(i_1, i_2, i_3) \in \mathbb{R}^3 | 0 \leq i_k \leq s-1, k \in [3]\}$, for $s \in [2, 10]$. The second and third columns in Table 1 shows the results of those experiments. Each entry reports the Stress value, given by the Frobenius norm between the original and obtained matrices.

A possible explanation for the significant difference between the Stress values obtained by the algorithms is the sensitivity of the spectral decomposition applied in the `isedm` routine. The `edmsph` algorithm is more stable in that respect because it solves a triangular linear system at each iteration. In order to improve the results of `isedm` we rewrote the algorithm using singular value decomposition instead of spectral decomposition. This "new" routine is called $\texttt{isedm}_2$ and its results on the Moré-Wu constructions are presented in last column of the Table 1. In all experiments reported in this paper all algorithms obtained the correct embedding dimension.

s	`edmsph`	`isedm`	$\texttt{isedm}_2$
6	4.2862×10^{-12}	$1.4213 \times 10^{+02}$	4.3257×10^{-12}
7	1.1311×10^{-11}	$2.2174 \times 10^{+02}$	1.0699×10^{-11}
8	2.6081×10^{-11}	$1.8635 \times 10^{+03}$	2.3190×10^{-11}
9	1.2382×10^{-10}	$1.9240 \times 10^{+03}$	1.8667×10^{-10}
10	1.2037×10^{-10}	$8.1544 \times 10^{+03}$	2.8984×10^{-10}

Table 1
Stress values obtained on Moré-Wu instances.

In the second series of experiments we generated 500 artificial molecular instances, with 1,000 points each, built according to the ideas of Lavor [4]. The Stress values obtained by `edmsph` were smaller than those obtained by $\texttt{isedm}_2$ in 412 of the datasets, amounting to 82.4% of the experiments.

Finally, in the third series of experiments we used 69 proteins from the Protein Data Bank with number of atoms ranging from 356 to 1,997. Stress values of `edmsph` were smaller than those of $\texttt{isedm}_2$ in 40 of the datasets.

4 Conclusions

We introduced an algorithm that decides whether a given symmetric hollow (i.e., with zero diagonal) matrix with nonnegative elements is a EDM. Additionally, if the matrix is indeed a EDM, the procedure computes the matrix's embedding dimension, alongside an actual embedding. Further work will focus on dealing with noisy distance data, and to the rank deficient case where the centers of K spheres in $\mathbb{R}^K$ do not span a $(K-1)$-dimensional subspace.

5 Acknowledgements

Financial support is acknowledged from the Brazilian research agencies CNPq and FAPESP, and from the French research agency ANR (project ANR-10-BINF-03-08 "Bip:Bip").

References

[1] Dattorro, J.,"Convex Optimization and Euclidean Distance Geometry", Meboo, Palo Alto, (2005).

[2] L. Liberti, Lavor, C., N. Maculan, and A. Mucherino, "Euclidean distance geometry and applications." SIAM Review 56 (2014): 3-69.

[3] Householder, A. S., and G. Young, *Discussion of a set of points in terms of their mutual distances.* Psychometrika 3 (1938): 19-22.

[4] Lavor, C., *On generating instances for the molecular distance geometry problem.* Global optimization. Springer US, (2006): 405-414.

[5] Moré, J. J., and Z. Wu, *Global continuation for distance geometry problems.* SIAM Journal on Optimization 7 (1997): 814-836.

[6] Dong, Q., and Z. Wu, *A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances.* Journal of Global Optimization 22 (2002): 365-375.

[7] Scheraga, H. A. and M. J. Sippl, *Solution of the embedding problem and decomposition of symmetric matrices.* Proceedings of the National Academy of Sciences 82 (1985): 2197-2201.

[8] Barvinok, A., *Problems of Distance Geometry and convex properties of quadratic maps.* Discrete and Computational Geometry 13 (1995): 189-202.