Euclidean Distance Matrices and the Molecular Conformation Problem

Abdo Y. Alfakih* and Henry Wolkowicz†

Department of combinatorics and optimization
University of Waterloo
Waterloo, Ontario Canada N2L 3G1

Report CORR 2002-17

July 16, 2002

1 Introduction

In molecular conformation theory [7] and in multidimensional scaling in statistics [17, 6], one is usually interested in the following problem known as the *graph realizability problem* (GRP). Given an edge-weighted undirected incomplete simple graph $G = (V, E, \omega)$ with node set $V = \{v_1, v_2, \ldots, v_n\}$, edge set $E \subset V \times V$, and weights $\omega : E \rightarrow \Re^+$, the GRP is the problem of determining whether or not G is realizable in some Euclidean space. A graph $G = (V, E, \omega)$ is said to be realizable in $\Re^r$ iff there exist points $p^1, p^2, \ldots, p^n$ in $\Re^r$ such that $\|p^i - p^j\|^2 = \omega_{ij}$ for all edges $(v_i, v_j) \in E$, where $\| \, . \, \|$ is the Euclidean norm.

In the context of molecular conformation theory, graph G represents a molecule with V, the set of nodes of G, representing the atoms. It is possible, using nuclear magnetic resonance [24, 22], to determine some of the pair-wise distances of the atoms. If the distance between atoms i and j is known, then edge (v_i, v_j) is created in graph G with weight ω_{ij} equal to the square of this distance. Naturally, the missing edges of G correspond to those unknown interatomic distances. Thus the shape of this molecule, which is important in determining its chemical and biological properties, can be found by solving the GRP. Of course in this case we want the graph to be realizable in $\Re^3$.

Euclidean distance matrices (EDMs) provide a very useful tool in the study of the GRP. In fact the GRP can be formulated as a Euclidean distance matrix completion problem [5]. Other optimization methods for the GRP are discussed in [18, 14, 21].

*E-mail: alfakih@alumni.engin.umich.edu

†E-mail: hwolkowi@orion2.math.uwaterloo.ca

An $n \times n$ matrix $D = (d_{ij})$ is said to be a *Euclidean distance matrix (EDM)* iff there exist n points $p^1, p^2, \ldots, p^n$ in some Euclidean space $\Re^r$, such that $\|p^i - p^j\|^2 = d_{ij}$ for all $i, j = 1, \ldots, n$. The dimension of the smallest Euclidean space containing points $p^1, p^2, \ldots, p^n$ is called the *embedding dimension* of D. Given graph $G = (V, E, \omega)$, define an $n \times n$ partial symmetric matrix $A_G = (a_{ij})$ with some entries specified (or fixed) and the rest unspecified (or free) such that

$$a_{ij} := \begin{cases} \omega_{ij} & \text{if and only if } (v_i, v_j) \in E, \\ \text{free} & \text{otherwise .} \end{cases} \quad (1)$$

An $n \times n$ matrix D_1 is said to be an EDM completion of A_G if and only if D_1 is EDM and $d_{1ij} = a_{ij}$ for all fixed elements of A_G. Thus it immediately follows that $G = (V, E, \omega)$ is realizable in $\Re^r$ if and only if A_G has an EDM completion of embedding dimension r.

In section 2 we present some of the properties of EDMs and we show how the GRP can be formulated as a semi-definite programming problem. In section 3, we study the problem of determining whether or not a given EDM completion is unique. Finally in section 4 we present an interesting connection between the set of EDMs and the set of correlation matrices.

2 Euclidean Distance Matrices

A necessary and sufficient condition for a symmetric matrix D with zero diagonal to be a EDM was given by Schoenberg [20]. The following is a slightly modified version of Schoenberg result due to Gower [12].

Let M be the subspace orthogonal to e, the vector of all ones, i.e,

$$M := \{e\}^{\perp} = \{x \in \Re^n : e^T x = 0\}.$$

Let J be the orthogonal projection on M, i.e., $J := I_n - ee^T/n$ where I_n is the identity matrix of order n. Then we have the following result.

Theorem 2.1 *A matrix D with zero diagonal is EDM if and only if the matrix JDJ is negative semidefinite. Furthermore, the embedding dimension of D is given by the rank of JDJ.*

Let D be a EDM and let rank $(JDJ) = r$. Then, the points $p^1, p^2, \ldots, p^n$ that generate D are given by the rows of the $n \times r$ matrix P where $-\frac{1}{2}JDJ := PP^T$. Note that the centroid of the points p^i, $i = 1, \ldots, n$ coincides with the origin. This follows since $P^T e = 0$ which is implied by the definition of J namely $Je = 0$.

An alternative statement of Theorem 2.1 which is more convenient for our purposes is given in the next theorem [8, 3]. Let $\mathcal{S}_n$ denote the space of symmetric matrices of order n and let $\mathcal{S}_H$ denote the subspace of $\mathcal{S}_n$ defined as

$$\mathcal{S}_H = \{B \in \mathcal{S}_n : \text{diag}\, B = 0\}.$$

Let V be the $n \times (n-1)$ matrix whose columns form an orthonormal basis of M; that is, V satisfies:

$$V^T e = 0 \,, \quad V^T V = I_{n-1} \,. \quad (2)$$

Note that $J = VV^T$. Now define the linear operators $\mathcal{K}_V : \mathcal{S}_{n-1} \to \mathcal{S}_H$ and $\mathcal{T}_V : \mathcal{S}_H \to \mathcal{S}_{n-1}$ such that

$$\begin{aligned} \mathcal{K}_V(X) = {} & \text{diag}\,(VXV^T)e^T + e(\text{diag}\,(VXV^T))^T \\ & - 2\,VXV^T, \\ \mathcal{T}_V(D) = {} & -\tfrac{1}{2}V^T DV \,. \end{aligned}$$

Theorem 2.2 *[3] The following statements are equivalent:*

1. *$D \in \mathcal{S}_H$ is a EDM.*

2. $\mathcal{T}_V(D)$ *is positive semidefinite.*

3. $D = \mathcal{K}_V(X)$ *for some positive semidefinite matrix* X *in* $\mathcal{S}_{n-1}$.

Let A_G be the partial matrix associated with $G = (V, E, \omega)$ Without any loss of generality we can set the free elements of A_G initially to zero. Let H be the adjacency matrix of graph G. Then graph G has a realization in $\Re^r$ for some positive integer $r \le n-1$ if and only if the optimal objective function of the following semidefinite programming problem is equal to zero.

$$\begin{array}{ll} \min & \mu(X) = ||H \circ (\mathcal{K}_V(X) - A_G)||_F^2 \\ \text{subject to} & \\ & X \succeq 0, \end{array} \tag{3}$$

where $||\ .\ ||_F$ is the Frobenius norm defined as $||A||_F^2 = \text{trace } A^T A$, $\circ$ denotes the Hadamard product, and $X \succeq 0$ means that the matrix X is symmetric positive semidefinite. If X^* is the optimal solution of problem (3) such that $\mu(X^*) = 0$, then $D_1 = \mathcal{K}_V(X^*)$ is an EDM completion of A_G. Note that we could answer the question whether graph G has a realization in $\Re^k$ for a given integer k simply by adding to Problem (3) the side constraint

$$\text{rank } X = k.$$

Unfortunately the feasible region of Problem (3) with this extra constraint is not convex. In fact, it was shown by Saxe [19] that given an edge-weighted graph G and an integer k, the problem of determining whether or not G is realizable in $\Re^k$ is NP-hard. Where as the GRP, i.e., the problem of determining whether or not G is realizable in some Euclidean space is still open [15], the GRP can be solved "approximately" in polynomial time since problem (3) can be solved using interior point algorithms for semidefinite programming [23]. For polynomial instances of the GRP see [16].

A primal-dual interior point algorithm for Problem (3) was given in [3] and numerical tests were also presented. Note that by taking H to be the matrix of all ones, problem (3) can be used to find the closest Euclidean distance matrix, in Frobenius sense, to a given matrix A. For another approach to solving the closest EDM problem see [1, 11].

3 Uniqueness of EDM Completions

Given a EDM completion D_1 of a given partial matrix A_G, an interesting problem is the problem of determining whether or not D_1 is unique. A characterization of the uniqueness of D_1 is discussed next using the notion of Gale transform from the theory of polytopes [10, 13].

Let $p^1, p^2, \ldots, p^n$ be points in $\Re^r$ whose centroid coincides with the origin. Assume that the points $p^1, p^2, \ldots, p^n$ are not contained in a proper hyperplane. Then the matrix

$$P := \begin{bmatrix} {p^1}^T \\ {p^2}^T \\ \vdots \\ {p^n}^T \end{bmatrix}$$

is of rank r. Let $\bar{r} = n - 1 - r$. For $\bar{r} \ge 1$, let Λ to be an $n \times \bar{r}$ matrix, whose columns form a basis for the null space of the $(r+1) \times n$ matrix $\begin{bmatrix} P^T \\ e^T \end{bmatrix}$; that is,

$$P^T \Lambda = 0, \quad e^T \Lambda = 0, \Lambda \text{ is full column rank.} \tag{4}$$

Λ is called a *Gale matrix* corresponding to the EDM matrix D generated by p^i, $i = 1, \ldots, n$; and the ith row of Λ, considered as a vector in $\Re^{\bar{r}}$, is called a Gale transform of p^i. It is clear that Λ is not unique. In fact, for any nonsingular $\bar{r} \times \bar{r}$ matrix Q, ΛQ is also a Gale matrix. We will exploit this property to define a special Gale matrix Z which is more convenient for our purposes.

Let us write Λ in block form as

$$\Lambda = \begin{bmatrix} \Lambda_1 \\ \Lambda_2 \end{bmatrix},$$

where Λ_1 is $\bar{r} \times \bar{r}$ and Λ_2 is $(r+1) \times \bar{r}$. Without loss of generality we can assume that Λ_1 is nonsingular. Then Z is defined as

$$Z := \Lambda {\Lambda_1}^{-1} = \begin{bmatrix} I_{\bar{r}} \\ \Lambda_2 {\Lambda_1}^{-1} \end{bmatrix}. \tag{5}$$

Let ${z^i}^T$ denote the i-th row of Z. i.e.,

$$Z := \begin{bmatrix} {z^1}^T \\ {z^2}^T \\ \vdots \\ {z^n}^T \end{bmatrix}.$$

Hence z^i, the Gale transform of p^i, for $i = 1, \ldots, \bar{r}$ is equal to the ith unit vector in $\Re^{\bar{r}}$. Now we have the following result.

Theorem 3.1 *[2] Let A be a given partial symmetric matrix and let D_1 be an EDM completion of A. Let $X_1 = \mathcal{T}_V(D_1)$ and let $\bar{r} = n - 1 - rank X_1$. Then*

1. *If $\bar{r} = 0$. Then D_1 is not unique.*
2. *If $\bar{r} = 1$, let Z be the Gale matrix corresponding to D_1 defined in (5). The following condition is necessary and sufficient for D_1 to be unique:*
   (a) There exists a positive definite $\bar{r} \times \bar{r}$ matrix Ψ such that ${z^i}^T \Psi z^j = 0$ for all i, j such that a_{ij} is free.

For a proof of this theorem and other results concerning the case $\bar{r} \geq 2$ see [2].

4 EDMs and the Elliptope

As it turned out, there is an interesting connection between the set of EDMs and the set of correlation matrices, i.e., the set of symmetric positive semi-definite matrices whose diagonal is equal to e. Denote the set of all $n \times n$ EDM matrices by $\mathcal{D}_n$. Then it easily follows from Theorem 2.1 that $\mathcal{D}_n$ is a closed convex cone. Let $\mathcal{E}_n$ denote the set of $n \times n$ correlation matrices often referred to as the *Elliptope* [9]. Next we discuss the relationship between $\mathcal{D}_n$ and $\mathcal{E}_n$.

Given a convex set $K \in \mathcal{S}_n$ and a point $A \in K$, let $R(K, A)$ and $N(K, A)$ denote, respectively, the *radial cone* and the *normal cone* of K at A; that is

$$R(K, A) = \{B : B = \lambda(X - A), \forall X \in K, \lambda \geq 0\},$$
$$N(K, A) = \{B : \langle B, A - X \rangle \geq 0, \forall X \in K\}.$$

Hence, $N(K, A) = (R(K, A))^\circ$, where Q° denotes the *polar* of cone Q; defined as

$$Q^\circ = \{B \in \mathcal{S}_n : \langle B, X \rangle \leq 0, \text{ for all } X \in Q\}.$$

Furthermore, $(N(K, A))^\circ =$ closure of $R(K, A) = T(K, A)$, where $T(K, A)$ is called the *tangent cone* of K at A. Let E denote the matrix of all ones. Then it is not difficult to show that

$$N(\mathcal{E}_n, E) = \{B : B = \operatorname{Diag} y - V \Phi V^T\},$$

where y is any vector in $\Re^n$ and Φ is any positive semidefinite matrix in $\mathcal{S}_{n-1}$. It would easily follows then that $\mathcal{D}_n = -\, T(\mathcal{E}_n, E)$; i.e., the cone of

EDMs of order n is equal to the negative of the tangent cone of the set of $n \times n$ correlation matrices at E, the matrix of all ones [9, Page 535].

An interesting question is whether the radial cone of $\mathcal{E}_n$ at E is closed. If this is the case then $R(\mathcal{E}_n, E) = T(\mathcal{E}_n, E)$ which would imply that for any EDM matrix D there exists a nonnegative scalar λ and a correlation matrix C such that $D = \lambda(E - C)$. Unfortunately this is not true. the radial cone $R(\mathcal{E}_n, E)$ is not closed as can be shown by the following example.

Let

$$D = \begin{bmatrix} 0 & 1 & 4 \\ 1 & 0 & 1 \\ 4 & 1 & 0 \end{bmatrix}.$$

Then clearly D is a EDM generated by the three points p^1, p^2 and p^3 on the x-axis of coordinates -1 , 0 and 1 respectively. It is also clear that there exists no $\lambda \geq 0$ such that $D = \lambda(E - C_1)$ for any correlation matrix C_1 since the matrix $\lambda E - D$ is indefinite for all λ's.

For a characterization of those EDM matrices D which can be expressed as $D = \lambda(E - C)$ for some scalar λ and some correlation matrix C see [4]. Here again this characterization is given in terms of the Gale transform of the points $p^1, p^2, \ldots, p^n$ that generate D.

References

[1] S. Al-Homidan and R. Fletcher. Hybrid methods for finding the nearest Euclidean distance matrix. *Recent advances in nonsmooth optmization*, pages 1–17, 1995.

[2] A. Y. Alfakih. On the uniqueness of Euclidean distance matrix completions. Technical report, 2002. submitted.

[3] A. Y. Alfakih, A. Khandani, and H. Wolkowicz. Solving Euclidean distance matrix completion problems via semidefinite programming. *Comput. Optim. Appl.*, 12:13–30, 1999.

[4] A. Y. Alfakih and H. Wolkowicz. Two theorems on Euclidean distance matrices and Gale transform. *Linear Algebra Appl.*, 340:149–154, 2002.

[5] M. Bakonyi and C. R. Johnson. The Euclidean distance matrix completion problem. *SIAM J. Matrix Anal. Appl.*, 16:646–654, 1995.

[6] I. Borg and P. Groenen. *Modern multidimensional scaling, theory and applications.* Springer - Verlag , New York, 1997.

[7] G. M. Crippen and T.F. Havel. *Distance Geometry and Molecular Conformation.* Wiley, New York, 1988.

[8] F. Critchley. On certain linear mappings between inner-product and squared distance matrices. *Linear Algebra Appl.*, 105:91–107, 1988.

[9] M. Deza and M. Laurent. *Geometry of Cuts and Metrics, Algorithms and Combinatorics. Vol 15.* Springer, Berlin, 1997.

[10] D. Gale. Neighboring vertices on a convex polyhedron. In *Linear inequalities and related system*, pages 255–263. Princeton University Press, 1956.

[11] W. Glunt, T. L. Hayden, S. Hong, and J. Wells. An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM J. Matrix Anal. Appl.*, 11:589–600, 1990.

[12] J. C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra Appl.*, 67:81–97, 1985.

[13] B. Grünbaum. *Convex polytopes.* John Wiley & Sons, 1967.

[14] B. Hendrickson. The molecule problem: Exploiting structure in global optimization. *SIAM J. Optim.*, 5:835–857, 1995.

[15] M. Laurent. Cuts, matrix completions and graph rigidity. *Mathematical Programming, Series B*, 79:255–283, 1997.

[16] M. Laurent. Polynomial instances of the positive semidefinite and Euclidean distance matrix completion problems. *SIAM J. Matrix Anal. Appl.*, 22:874–894, 2000.

[17] J. De Leeuw and W. Heiser. Theory of multidimensional scaling. In P. R. Krishnaiah and L. N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 285–316. North-Holland, 1982.

[18] J. Moré and Z. Wu. Distance geometry optimization for protein structures. *J. Global Optim.*, 15:219–234, 1999.

[19] J. B. Saxe. Embeddability of weighted graphs in k-space is strongly NP-hard. *Proc. 17th Allerton Conf. in Communications, Control, and Computing*, pages 480–489, 1979.

[20] I.J. Schoenberg. Remarks to Maurice Frechet's article: Sur la definition axiomatique d'une classe d'espaces vectoriels distancies applicables vectoriellement sur l'espace de Hilbert. *Ann. Math.*, 36:724–732, 1935.

[21] M. W. Trosset. Distance matrix completion by numerical optimization. *Comput. Optim. Appl.*, 17:11–22, 2000.

[22] C. P. Wells. *An improved method for sampling of molecular conformation space.* PhD thesis, University of Kentucky, 1995.

[23] H. Wolkowicz, R. Saigal, and L. Vandenberghe, editors. *Handbook of Semidefinite Programming. Theory, Algorithms and Applications.* Kluwer Academic Publishers, Boston MA, 2000.

[24] K. Wütrich. The development of nuclear magnetic resonance spectroscopy as a technique for protein structure determination. *Accounts of chemical research*, 22:36–44, 1989.