Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2197–2201, April 1985
Applied Physical and Mathematical Sciences

Solution of the embedding problem and decomposition of symmetric matrices

(distance geometry/embedding of distances/matrix factorization/Cholesky factorization/quadratic forms)

MANFRED J. SIPPL*†‡ AND HAROLD A. SCHERAGA*§

*Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853; and †Institute for General Biology, Biochemistry and Biophysics, University of Salzburg, Erzabt Klotz Strasse 11, A-5020 Salzburg, Austria

Contributed by Harold A. Scheraga, December 3, 1984

ABSTRACT **A solution of the problem of calculating cartesian coordinates from a matrix of interpoint distances (the embedding problem) is reported. An efficient and numerically stable algorithm for the transformation of distances to coordinates is then obtained. It is shown that the embedding problem is intimately related to the theory of symmetric matrices, since every symmetric matrix is related to a general distance matrix by a one-to-one transformation. Embedding of a distance matrix yields a decomposition of the associated symmetric matrix in the form of a sum over outer products of a linear independent system of coordinate vectors. It is shown that such a decomposition exists for every symmetric matrix and that it is numerically stable. From this decomposition, the rank and the numbers of positive, negative, and zero eigenvalues of the symmetric matrix are obtained directly.**

The embedding problem can be stated as follows: Given a complete set of distances between points of known connexity, calculate cartesian coordinates for the associated set of points. A solution to the embedding problem would also yield the dimension—i.e., the number of coordinate axes—occupied by the system of points. In addition to theoretical interest in the embedding problem, procedures for its solution have important applications in various fields of applied mathematics. In the field of molecular structure, for example, a number of experimental procedures yield interatomic distances as primary data.

The origin of distance geometry dates back to the work of Menger (1), who coined the term metric geometry. He proposed that geometry should be studied in terms of distances, in addition to the more traditional approaches of axiomatic and analytical geometry. This branch of geometry rapidly expanded, especially as a consequence of the work of Blumenthal (2).

A useful theorem in distance geometry was proven by Schoenberg (3):

THEOREM 1: *A necessary and sufficient condition that the matrix* $\mathbf{D} = \{d_{ij}\}$ *represents the distances of a system of* $N + 1$ *points* $P_0, P_1, \ldots, P_N$ *in euclidian space* E^M *but not in* E^{M-1} *is that the quadratic form*

$$F(x_1, x_2, \ldots, x_N) = \frac{1}{2}\sum_{i,j=1}^{N}(d_{0i}^2 + d_{0j}^2 - d_{ij}^2)x_i x_j = \sum_{i,j=1}^{N} v_{ij}\, x_i x_j = \mathbf{x}^T\mathbf{V}\mathbf{x} \qquad [1]$$

be positive and of rank M, *where* $\mathbf{V} = \{v_{ij}\}$ *is a symmetric matrix, with* $v_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$. The special point P_0 is at the origin. All vectors are column vectors unless their transpose is indicated.

Schoenberg showed that the actual construction of the coordinates of the $N + 1$ points is equivalent to reducing F to its canonical form

$$F(x_1, x_2, \ldots, x_N) = \sum_{\mu=1}^{M} y_\mu^2. \qquad [2]$$

Then the elements of an $M \times N$ matrix $\mathbf{C}$ such that

$$\mathbf{y} = \mathbf{C}\,\mathbf{x} \qquad [3]$$

represent the coordinates of the points $P_1, P_2, \ldots, P_N$—i.e., $P_j = (c_{1j}, c_{2j}, \ldots, c_{Mj})$ is identical to the jth column of $\mathbf{C}$, and $P_0 = (0, 0, \ldots, 0)$ is at the origin. The vectors $\mathbf{c}_\mu^T = (c_{\mu 1}, c_{\mu 2}, \ldots, c_{\mu N})$ formed by the rows of $\mathbf{C}$ represent the coordinates of all points along the dimension μ.

Schoenberg also investigated the case in which F is indefinite. In this case F can always be reduced to

$$F(x_1, x_2, \ldots, x_N) = \sum_{\mu=1}^{M} \varepsilon_\mu y_\mu^2 \qquad \varepsilon_\mu = \pm 1. \qquad [4]$$

The points $P_0, P_1, \ldots, P_N$ are then embeddable in a pseudo-euclidian space ψ_{pq} with p real and q imaginary dimensions, where p and q are equal to the numbers of positive and negative ε's, respectively, in Eq. **4** (3). The case of positive F is then the special one with $q = 0$. In the pseudo-euclidian space, the calculation of coordinates is again reduced to the problem of finding a linear transformation $\mathbf{y} = \mathbf{C}\,\mathbf{x}$ such that F is in the canonical form (Eq. **4**).

Such a transformation $\mathbf{C}$ is obtained with the help of *Lemma 1*.

LEMMA 1: $F(x_1, x_2, \ldots, x_N)$ *is reduced to its canonical form* (Eq. **4**) *if the matrix* $\mathbf{V}$ *of coefficients* v_{ij} *of* F *is decomposed to the outer product form*

$$\mathbf{V} = \sum_{\mu=1}^{M} \varepsilon_\mu \mathbf{c}_\mu \mathbf{c}_\mu^T. \qquad [5]$$

Making use of Eqs. **1** and **5**, the proof is

$$F(\mathbf{x}) = \mathbf{x}^T \sum_{\mu=1}^{M} \varepsilon_\mu \mathbf{c}_\mu \mathbf{c}_\mu^T \mathbf{x} = \sum_{\mu=1}^{M} \varepsilon_\mu (\mathbf{c}_\mu^T \mathbf{x})^2 = \sum_{\mu=1}^{M} \varepsilon_\mu y_\mu^2. \qquad [6]$$

From Eq. **6**, we immediately find the transformation $\mathbf{C}$, since the rows of $\mathbf{C}$ are the $(\varepsilon_\mu)^{1/2}\,\mathbf{c}_\mu^T$ in Eq. **6**.

In this paper, we present an algorithm for the calculation of coordinates for the general pseudo-euclidian embedding problem and for the decomposition of $\mathbf{V}$ in the form of Eq. **5**. The algorithm will decompose a matrix $\mathbf{D}^2$ of squared distances d_{ij}^2 to the form

$$d_{ij}^2 = \sum_{\mu=1}^{M} \varepsilon_\mu (c_{i\mu} - c_{j\mu})^2 = \sum_{\mu=1}^{p} (c_{i\mu} - c_{j\mu})^2 - \sum_{\mu=p+1}^{M} (c_{i\mu} - c_{j\mu})^2, \qquad [7]$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

‡Present address: Baker Laboratory of Chemistry, Cornell University, Ithaca, NY 14853.
§To whom reprint requests should be addressed.

where ψ_{pq} is the space of smallest dimension in which the points $P_0, P_1, \ldots, P_N$ are embeddable and the $c_{i\mu}$ are their cartesian coordinates.

Embedding of $\mathbf{D}^2$ immediately yields the decomposition of V in the form of Eq. 5. Since any decomposition (Eq. 5) of V is equivalent to a factorization of V in the form

$$\mathbf{V} = \sum_{\mu=1}^{M} \varepsilon_\mu \mathbf{c}_\mu \mathbf{c}_\mu^T = \mathbf{C}^T\mathbf{C}, \qquad [8]$$

where $v_{ij} = \Sigma_\mu\ \varepsilon_\mu c_{i\mu} c_{j\mu}$ and the rows of C are the vectors $(\varepsilon_\mu)^{1/2}\mathbf{c}_\mu^T$, we can factor every symmetric matrix by embedding. If V is the matrix of coefficients of Schoenberg's quadratic form F, then the embedding algorithm yields a canonical transformation $\mathbf{y} = \mathbf{C}\,\mathbf{x}$. In most cases, we are able to obtain C in lower triangular form. In fact, a variant of the embedding algorithm is identical to the Cholesky factorization of V. The Cholesky factorization does not exist if the diagonal elements of V are zero, or it is unstable if these elements are very small. We shall show that we can always obtain a stable factorization of V in the form of Eq. 8 by embedding, although the factors are generally not triangular.

The Embedding Procedure

We now develop an algorithm for the embedding of a distance matrix whose associated quadratic form (Eq. 1) is positive definite. The solution for the pseudo-euclidian problem will then follow from a generalization of the euclidian embedding.

From Schoenberg's theorem 1, it follows that every $(N + 1) \times (N + 1)$ distance matrix with positive quadratic form F, with matrix of coefficients $\mathbf{V} = \{v_{ij}\}$, corresponds to an $N + 1$ simplex with vertices $P_0, P_1, \ldots, P_N$ and edges $d_{ij} = |P_i - P_j|$, which is embeddable in euclidian space $\mathbf{E}^M$ where $M \leq N$. Therefore, there exists a representation of $\mathbf{D}^2$ (in fact infinitely many) of the form

$$d_{ij}^2 = \sum_{\mu=1}^{M} (c_{i\mu} - c_{j\mu})^2 \qquad i, j = 0, 1, \ldots, N \qquad [9]$$

and we shall now derive such a decomposition (i.e., Eq. 9).

The three distances d_{ij}, d_{ik}, and d_{jk} define a triangle with vertices P_i, P_j, and P_k (Fig. 1A). These distances are related by the cosine law

$$\cos\beta_i = (d_{ij}^2 + d_{ik}^2 - d_{jk}^2)/2d_{ij}d_{ik}. \qquad [10]$$

We therefore are able to calculate the projection $\bar{d}_{ij}$ of d_{ij} on the axis defined by d_{ik}

$$\bar{d}_{ij} = d_{ij}(\cos\beta_i) = (d_{ij}^2 + d_{ik}^2 - d_{jk}^2)/2d_{ik}. \qquad [11]$$

Since

$$\cos\beta_i \begin{cases} < 0 \text{ for } \beta_i > \pi/2 \\ = 0 \quad \beta_i = \pi/2 \\ > 0 \quad \beta_i < \pi/2 \end{cases} \quad \text{with } (0 \leq \beta_i \leq \pi),$$

the projection $\bar{d}_{ij}$ on d_{ik} has the sign of $\cos\beta_i$ and therefore has magnitude and direction. Forming all possible triangles with the distance d_{ik}, we calculate the projections $\bar{d}_{ij}$ of all distances d_{ij} for $j = 0, 1, 2, \ldots, N$ on d_{ik}. Specifically,

$$\bar{d}_{ii} = (d_{ii}^2 + d_{ik}^2 - d_{ik}^2)/2d_{ik} = 0 \qquad [12]$$

and

$$\bar{d}_{ik} = (d_{ik}^2 + d_{ik}^2 - d_{kk}^2)/2d_{ik} = d_{ik}. \qquad [13]$$

As shown in Fig. 1B, the projections of all remaining distances d_{jl} $(l, j \neq i, k)$ on d_{ik} are obtained by

$$\bar{d}_{jl}^2 = (\bar{d}_{il} - \bar{d}_{ij})^2. \qquad [14]$$

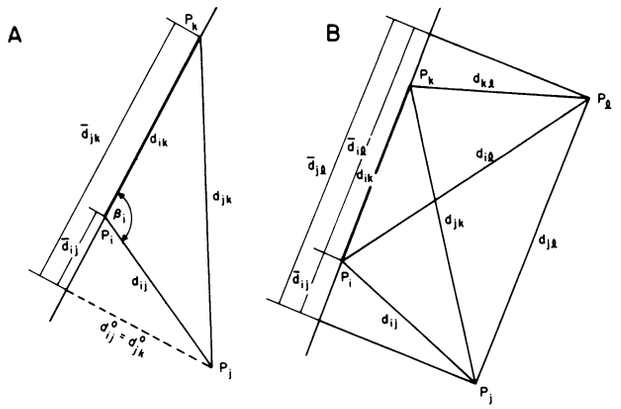

FIG. 1. Geometric relationships involved in the resolution of distances into parallel and orthogonal components relative to a generating distance d_{ik}. (*A*) Resolution of distances d_{ij} and d_{jk} into parallel components $\bar{d}_{ij}$ and $\bar{d}_{jk}$, respectively, and orthogonal components $d^o_{ij} = d^o_{jk}$, along the generating distance d_{ik}. See Eq. 11. (*B*) Projection of distance d_{jl} (which is not directly connected to the generating distance d_{ik}) on d_{ik}. It should be noted that the four points P_i, P_j, P_k, and P_l generally are not in a plane. See Eqs. 14 and 15.

Eq. 14 defines the parallel components of any distance d_{jl} along d_{ik}. The orthogonal component d^o_{jl} of d_{jl} relative to d_{ik} is defined by

$$(d^o_{jl})^2 = d_{jl}^2 - (\bar{d}_{jl})^2 = d_{jl}^2 - (\bar{d}_{il} - \bar{d}_{ij})^2. \qquad [15]$$

Comparison of Eqs. 9 and 15 yields

$$\begin{aligned} d_{jl}^2 &= (\bar{d}_{il} - \bar{d}_{ij})^2 + (d^o_{jl})^2 \\ &= (c_{l1} - c_{j1})^2 + \sum_{\mu=2}^{M} (c_{l\mu} - c_{j\mu})^2, \end{aligned} \qquad [16]$$

and it follows that Eq. 16 is satisfied if we set

$$c_{j1} = \bar{d}_{ij} \qquad j = 0, 1, 2, \ldots, N. \qquad [17]$$

We therefore obtain the vector $\mathbf{c}_1 = (c_{01}, c_{11}, c_{21}, \ldots, c_{N1})$ of cartesian coordinates of the $N + 1$ points along the axis defined by d_{ik}. Note that we can choose any $d_{ik} \neq 0$ to generate a vector $\mathbf{c}_1$. For further reference, we use the term generating distance for d_{ik}. From Eq. 11, it follows that only columns (or rows) i and k of the matrix D are needed to calculate $\mathbf{c}_1$.

Obviously, the matrix with the orthogonal components d^o_{jl} $(j, l = 0, 1, \ldots, N)$ is either a distance matrix or all of its elements are zero. In the latter case, the system of points is embeddable along d_{ik}—i.e., in a space of dimension $M = 1$—and the embedding is completed. Otherwise, we can treat the orthogonal matrix as described above.

We observe that $d^o_{ik} = 0$. Geometrically, this means that the points i and k have fused upon the projection along d_{ik} so that $d^o_{ij} = d^o_{jk}$ $(j = 0, 1, 2, \ldots, N)$ (Fig. 1A); i.e., the rows (and columns) i and k of $\mathbf{D}^o$ are identical. This follows from

$$\begin{aligned} (d^o_{jk})^2 &= d_{jk}^2 - (\bar{d}_{jk})^2 = d_{jk}^2 - (\bar{d}_{ik} - \bar{d}_{ij})^2 \\ &= d_{jk}^2 - \bar{d}_{ij}^2 - d^2_{ik} + 2\bar{d}_{ij}d_{ik} \end{aligned} \qquad [18]$$

since $\bar{d}_{ik} = d_{ik}$. Insertion of Eq. 11 for d_{ij} then yields

$$\begin{aligned} (d^o_{jk})^2 &= d_{jk}^2 - \bar{d}_{ij}^2 - d_{ik}^2 + d_{ij}^2 + d_{ik}^2 - d_{jk}^2 \\ &= d_{ij}^2 - \bar{d}_{ij}^2 = (d^o_{ij})^2. \end{aligned} \qquad [19]$$

We use the following notation: $d_{ij} = d_{ij}(1)$, $\bar{d}_{ij} = \bar{d}_{ij}(1)$, and $d^o_{ij} = d_{ij}(2)$, where the integer in parentheses refers to the number of projections. $\mathbf{D}(1) = \{d_{ij}(1)\}$ is the original distance matrix D, used for the first projection. $\mathbf{D}(2) = \{d^o_{ij}\} = \{d_{ij}(2)\}$ is used in the second projection, and so on.

The resolution of the distances $d_{ij}(2)$ into orthogonal and parallel components along a generating distance $d_{ik}(2)$ (which must not be zero) yields a vector of coordinates $\mathbf{c}_2$ and the $(N-1)\times(N-1)$ matrix $\mathbf{D}(3)$. The process can be repeated until all elements of the matrix $\mathbf{D}(M+1)$ are zero. This yields a solution of the embedding problem in terms of M vectors $\mathbf{c}_1, \ldots, \mathbf{c}_M$ and the dimension M of the space. The elements $d_{ij}(1)$ are now expressed as

$$d_{ij}^2(1) = \sum_{\mu=1}^{M} (c_{i\mu} - c_{j\mu})^2 \qquad i,j = 0, 1, \ldots, N. \qquad [20]$$

Eq. **20** thus provides cartesian coordinates $c_{i\mu}$ and $c_{j\mu}$ from the distances $d_{ij}^2(1)$.

Embedding in Pseudo-euclidian Spaces

In the euclidian problem the ε_μ's in

$$d_{ij}^2 = \sum_{\mu=1}^{M} \varepsilon_\mu (c_{i\mu} - c_{j\mu})^2 \qquad \varepsilon_\mu = \pm 1 \qquad [21]$$

are all positive, and hence the squared distances d_{ij}^2 are also positive. If some or all of the ε_μs are equal to -1, then d_{ij}^2 may be or has to be negative and the distances d_{ij} themselves are imaginary. Therefore, if we take a negative $d_{ik}^2(m)$ as the mth generating distance, the coordinates obtained from Eqs. **11** and **17** will be imaginary. The generalization of Eqs. **11** and **17** to the pseudo-euclidian problem therefore leads to

$$\begin{aligned} c_{jm} &= (\varepsilon_m)^{1/2}\overline{d}_{ij}(m) \\ &= [d_{ij}^2(m) + d_{ik}^2(m) - d_{jk}^2(m)]/2(|d_{ik}^2(m)|)^{1/2}, \end{aligned} \qquad [22]$$

where $\varepsilon_m = \text{sign}[d_{ik}^2(m)]$ and m is the projection number (i.e., $m = 1, 2, \ldots$ in the first, second, etc., projection); similarly Eqs. **15** and **17** lead to

$$d_{ij}^2(m+1) = d_{ij}^2(m) - \varepsilon_m (c_{im} - c_{jm})^2. \qquad [23]$$

Eqs. **22** and **23** yield a decomposition of any matrix $\mathbf{D}^2$ whose elements satisfy $d_{ii}^2 = 0$ and $d_{ij}^2 = d_{ji}^2$, $i, j = 0, 1, \ldots, N$. Therefore, every such matrix is embeddable in ψ_{pq} with $p + q = M \le N$. The coordinate vectors, $\mathbf{c}_\mu$ ($\mu = 1, 2, \ldots, M$), form a basis set of ψ_{pq} that is generally not orthogonal.

Embedding is not unique because different choices of generating distances yield different sets of coordinates. However, we shall show that the vectors $\mathbf{c}_\mu$ obtained from the embedding algorithm (Eqs. **22** and **23** are linearly independent so that the dimension M and the numbers p and q of real and imaginary axes, respectively, are invariants of the embedding.

THEOREM 2: *The vectors* $\mathbf{c}_\mu = (c_{0\mu}, c_{1\mu}, \ldots, c_{N\mu})$, $\mu = 1, 2, \ldots, M$, *obtained from the embedding algorithm are linearly independent.*

To show this we need *Lemma 2.*

LEMMA 2: *Let* $d_{ik}(m)$ *be the generating distance in step* m. *Then* $c_{i\mu} = c_{k\mu}$, $\mu > m$. From Eq. **19**, which also holds for the pseudo-euclidian case, we have, after including the step parameter m,

$$\begin{aligned} d_{jk}^2(m+1) = d_{kj}^2(m+1) = d_{ij}^2(m+1) \\ j = 0, 1, \ldots, N, \end{aligned} \qquad [24]$$

where $d_{ik}(m)$ is the generator in step m. Let $d_{st}(m+1)$ be the generating distance in step $m+1$. Then, from Eq. **22**,

$$\begin{aligned} c_{i,m+1} &= (\varepsilon_{m+1})^{1/2}\,\overline{d}_{si}(m+1) \\ &= \frac{[d_{si}^2(m+1) + d_{st}^2(m+1) - d_{ti}^2(m+1)]}{2(|d_{st}(m+1)|)^{1/2}} \end{aligned} \qquad [25]$$

and

$$\begin{aligned} c_{k,m+1} &= (\varepsilon_{m+1})^{1/2}\,\overline{d}_{sk}(m+1) \\ &= \frac{[d_{sk}^2(m+1) + d_{st}^2(m+1) - d_{tk}^2(m+1)]}{2(|d_{st}(m+1)|)^{1/2}} \end{aligned}$$

so that, from Eq. **24**, $c_{i,m+1} = c_{k,m+1}$. Since (from Eqs. **23** and **24**)

$$\begin{aligned} d_{ij}^2(m+2) &= d_{ij}^2(m+1) - \varepsilon_{m+1}(c_{i,m+1} - c_{j,m+1})^2 \\ &= d_{kj}^2(m+1) - \varepsilon_{m+1}(c_{k,m+1} - c_{j,m+1})^2 \\ &= d_{kj}^2(m+2) \end{aligned} \qquad [26]$$

for $j = 0, 1, \ldots, N$, we have $c_{i,m+2} = c_{k,m+2}$ and generally

$$c_{i\mu} = c_{k\mu} \qquad \mu = m+1, m+2, \ldots, M$$

and

$$\begin{aligned} d_{ij}^2(\mu) = d_{kj}^2(\mu) \\ \mu = m+1, m+2, \ldots, M; j = 0, 1, \ldots, N. \end{aligned} \qquad [27]$$

We can now prove *Theorem 2.* For this purpose we assume the opposite—i.e., that $\mathbf{c}_1$ can be expressed as a linear combination

$$\mathbf{c}_1 = \sum_{\mu=2}^{M} a_\mu \mathbf{c}_\mu. \qquad [28]$$

Then $\mathbf{c}_1$ is generated by $d_{ik}(1)$ and, from Eqs. **13**, **21**, and **22**,

$$\begin{aligned} d_{ik}^2(1) = \varepsilon_1\,\overline{d}_{ik}^2(1) = \varepsilon_1\left[\sum_{\mu=2}^{M} a_\mu(c_{i\mu} - c_{k\mu})\right]^2 \\ + \sum_{\mu=2}^{M} \varepsilon_\mu (c_{i\mu} - c_{k\mu})^2. \end{aligned} \qquad [29]$$

But $c_{i\mu} = c_{k\mu}$, $\mu = 2, 3, \ldots, M$, from *Lemma 2*, so that $d_{ik}(1) = 0$, in contradiction to our assumption that $d_{ik}(1)$ is the generator of $\mathbf{c}_1$. This also proves *Theorem 3.*

THEOREM 3: *Different sequences of generating distances yield the same number* M *of independent coordinate vectors.* Otherwise, the coordinate vectors $\mathbf{c}_\mu$ would be linearly dependent. Also, from the linear independence, M is the smallest possible number of such vectors.

To show that the nature of the space ψ_{pq} is invariant under embedding, we prove *Theorem 4.*

THEOREM 4: *The numbers* p *and* q *of real and imaginary axes are invariants of the embedding procedure—i.e., they are independent of the specific sequence of generators.*

Assume that two different decompositions yield

$$\begin{aligned} d_{ij}^2 &= \sum_{\mu=1}^{p} (c_{i\mu} - c_{j\mu})^2 - \sum_{\mu=p+1}^{M} (c_{i\mu} - c_{j\mu})^2 \\ &= \sum_{\mu=1}^{r} (b_{i\mu} - b_{j\mu})^2 - \sum_{\mu=r+1}^{M} (b_{i\mu} - b_{j\mu})^2, \end{aligned} \qquad [30]$$

where the b's are a second set of coordinates. From Eq. **1**, we form the matrix $v_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2$. From Eq. **30**, this becomes

$$\begin{aligned} v_{ij} &= \sum_{\mu=1}^{p} (c_{0\mu}^2 - c_{i\mu}c_{0\mu} - c_{j\mu}c_{0\mu} + c_{i\mu}c_{j\mu}) \\ &\quad - \sum_{\mu=p+1}^{M} (c_{0\mu}^2 - c_{i\mu}c_{0\mu} - c_{j\mu}c_{0\mu} + c_{i\mu}c_{j\mu}) \\ &= \sum_{\mu=1}^{r} (b_{0\mu}^2 - b_{i\mu}b_{0\mu} - b_{j\mu}b_{0\mu} + b_{i\mu}b_{j\mu}) \\ &\quad - \sum_{\mu=r+1}^{M} (b_{0\mu}^2 - b_{i\mu}b_{0\mu} - b_{j\mu}b_{0\mu} + b_{i\mu}b_{j\mu}). \end{aligned} \qquad [31]$$

Set $\mathbf{s}_\mu = (s_{1\mu}, s_{2\mu}, \ldots, s_{N\mu})$ where $s_{i\mu} = (c_{0\mu} - c_{i\mu})$ and $\mathbf{t}_\mu = (t_{1\mu}, t_{2\mu}, \ldots, t_{N\mu})$ where $t_{i\mu} = (b_{0\mu} - b_{i\mu})$. Then, from Eq. **31**,

$$v_{ij} = \sum_{\mu=1}^{p} s_{i\mu}s_{j\mu} - \sum_{\mu=p+1}^{M} s_{i\mu}s_{j\mu} = \sum_{\mu=1}^{r} t_{i\mu}t_{j\mu} - \sum_{\mu=r+1}^{M} t_{i\mu}t_{j\mu} \qquad [32]$$

and, from Eq. **8**,

$$\mathbf{V} = \sum_{\mu=1}^{p} \mathbf{s}_\mu \mathbf{s}_\mu^T - \sum_{\mu=p+1}^{M} \mathbf{s}_\mu \mathbf{s}_\mu^T = \sum_{\mu=1}^{r} \mathbf{t}_\mu \mathbf{t}_\mu^T - \sum_{\mu=r+1}^{M} \mathbf{t}_\mu \mathbf{t}_\mu^T. \quad [33]$$

Hence, the quadratic form $F(\mathbf{x})$, with $\mathbf{V}$ as coefficient matrix, is in its canonical form since it follows from *Lemma 1* that

$$\begin{aligned}\mathbf{x}^T\mathbf{V}\mathbf{x} &= \mathbf{x}^T \sum_{\mu=1}^{p} \mathbf{s}_\mu \mathbf{s}_\mu^T \mathbf{x} - \mathbf{x}^T \sum_{\mu=p+1}^{M} \mathbf{s}_\mu \mathbf{s}_\mu^T \mathbf{x} \\ &= \sum_{\mu=1}^{p} (\mathbf{x}^T \mathbf{s}_\mu)^2 - \sum_{\mu=p+1}^{M} (\mathbf{x}^T \mathbf{s}_\mu)^2 \\ &= \sum_{\mu=1}^{r} (\mathbf{x}^T \mathbf{t}_\mu)^2 - \sum_{\mu=p+1}^{M} (\mathbf{x}^T \mathbf{t}_\mu)^2. \quad [34]\end{aligned}$$

From Sylvester's law of inertia (see, for example, ref. 4), the numbers of positive and negative squares in the canonical representation (Eq. **34**) of a quadratic form are invariant; i.e., two different canonical forms of $F(\mathbf{x})$ necessarily have the same numbers of positive and negative squares. Hence, $p = r$; i.e., any decomposition of $\mathbf{D}^2$ necessarily yields the same numbers p and q of real and imaginary dimensions of ψ_{pq}.

Decomposition of Symmetric Matrices

The embedding algorithm yields a decomposition of $\mathbf{D}^2$:

$$\begin{aligned}d_{ij}^2 &= \sum_{\mu=1}^{M} \varepsilon_\mu (c_{i\mu} - c_{j\mu})^2 \\ &= \sum_{\mu=1}^{M} \varepsilon_\mu c_{i\mu}^2 + \sum_{\mu=1}^{M} \varepsilon_\mu c_{j\mu}^2 - 2 \sum_{\mu=1}^{M} \varepsilon_\mu c_{i\mu} c_{j\mu}. \quad [35]\end{aligned}$$

Each term on the right-hand side of Eq. **35** is of the form

$$v_{ij} = \sum_{\mu=1}^{M} \varepsilon_\mu c_{i\mu} c_{j\mu}, \quad [36]$$

which may be regarded as the elements of a symmetric matrix $\mathbf{V}$; i.e., we have started with a matrix $\mathbf{D}^2$ and obtained the symmetric matrix $\mathbf{V}$. It is tempting to try to reverse the process, to obtain a decomposition of any symmetric matrix $\mathbf{U}$ in the form of Eq. **36** by transforming $\mathbf{U}$ to the associated matrix of squared distances $\mathbf{D}^2(\mathbf{U})$ with

$$d_{ij}^2 = u_{ii} + u_{jj} - 2u_{ij} \quad [37]$$

and applying the embedding algorithm to $\mathbf{D}^2(\mathbf{U})$; u_{ij} is the analog of v_{ij} of Eq. **36**.

Eqs. **1** and **37** lead to the transformation

$$u_{ij} = (d_{0i}^2 + d_{0j}^2 - d_{ij}^2)/2, \quad [38]$$

in which it can be seen that u_{0j} is equal to zero for $j = 0, 1, \ldots, N$. From this, we see that, if the transformation of an arbitrary symmetric $N \times N$ matrix $\mathbf{U}$ with elements u_{ij} $(i, j = 1, 2, \ldots, N)$ to its associated distance matrix $\mathbf{D}^2(\mathbf{U})$ is to be invertible in a unique way, we must border $\mathbf{U}$ by a *0*th row and *0*th column of zeros. Then the elements $d_{0j}^2 = u_{00} + u_{jj} - 2u_{0j} = u_{jj}$ $(j = 0, 1, \ldots, N)$ are the diagonal elements of $\mathbf{U}$. Their geometric meaning is that they represent the squared distances of the points $P_1, \ldots, P_N$ (associated with $\mathbf{U}$) to the origin P_0. Hence, for every symmetric matrix, we have a unique origin defined by the diagonal elements of $\mathbf{U}$.

The decomposition of $\mathbf{D}^2(\mathbf{U})$ then yields the coordinates $\bar{c}_{0\mu}$ of the origin of $\mathbf{U}$. But, generally, $\bar{c}_{0\mu} \neq 0$; i.e., the system represented by $\mathbf{U}$ generally will be translated by the embedding procedure so that

$$\bar{u}_{0j} = \sum_{\mu=1}^{M} \varepsilon_\mu \bar{c}_{0\mu} \bar{c}_{j\mu} \neq 0 \qquad (j = 0, 1, \ldots, N) \quad [39]$$

and $\bar{\mathbf{U}} \neq \mathbf{U}$. We then obtain a decomposition of $\mathbf{U}$ by translating the coordinates so that

$$c_{i\mu} = \bar{c}_{i\mu} - \bar{c}_{0\mu} \qquad (\mu = 1, 2, \ldots, M) \quad [40]$$

and $\mathbf{U}$ is decomposed to $\mathbf{U} = \sum_{\mu=1}^{M} \varepsilon_\mu \mathbf{c}_\mu \mathbf{c}_\mu^T$ or, equivalently,

$$u_{ij} = \sum_{\mu=1}^{M} \varepsilon_\mu c_{i\mu} c_{j\mu} \qquad (i, j = 1, 2, \ldots, N). \quad [41]$$

The embedding procedure therefore yields a decomposition of every symmetric matrix $\mathbf{U}$.

Several additional points follow from this result. First, consider the eigenvalue decomposition of $\mathbf{U}$ and Eq. **41**:

$$\mathbf{U} = \sum_{\mu=1}^{M} \lambda_\mu \mathbf{x}_\mu \mathbf{x}_\mu^T = \sum_{\mu=1}^{M} \varepsilon_\mu \mathbf{y}_\mu \mathbf{y}_\mu^T = \sum_{\mu=1}^{M} \varepsilon_\mu \mathbf{c}_\mu \mathbf{c}_\mu^T, \quad [42]$$

where the λ_μ's are the nonzero eigenvalues of $\mathbf{U}$ with eigenvectors $\mathbf{x}_\mu$, $\mathbf{y}_\mu = (|\lambda_\mu|)^{1/2}\mathbf{x}_\mu$, and $\varepsilon_\mu = \text{sign}(\lambda_\mu)$. Then, from Sylvester's law of inertia (4), we have *Corollary 1*.

COROLLARY 1: *The numbers of positive and negative eigenvalues of* $\mathbf{U}$ *are equal to the numbers* p *and* q *of* ψ_{pq} *obtained from embedding of* $\mathbf{D}^2(\mathbf{U})$, $M - (p + q)$ *is the number of zero eigenvalues of* $\mathbf{U}$, *and* $p + q$ *is the rank of* $\mathbf{U}$.

As a second point, Eq. **8** indicates that the decomposition of $\mathbf{U}$ (Eq. **41**) is equivalent to a factorization of $\mathbf{U}$:

$$\mathbf{U} = \mathbf{C}^T\mathbf{C}. \quad [43]$$

If $\mathbf{U}$ is a nonsingular matrix in the system of linear equations, $\mathbf{U}\,\mathbf{x} = \mathbf{y}$, a well-known procedure to solve for $\mathbf{x}$ in terms of $\mathbf{y}$ is to factor $\mathbf{U}$ so that the factors in Eq. **43** are upper and lower triangular. Under certain circumstances, embodied in *Theorem 5*, embedding can produce triangular factors.

THEOREM 5: *If the sequence of generating distances in the embedding of* $\mathbf{D}^2(\mathbf{U})$ *is* $d_{0,1}(1), d_{1,2}(2), \ldots, d_{m-1,m}(m), \ldots$ *then the factor* $\mathbf{C}$ *is upper triangular*; i.e., $c_{i\mu} = 0$, $i < \mu$.

From Eq. **22**, we have $c_{m-1,m} = (\varepsilon_m)^{1/2}\, \bar{d}_{m-1,m-1}(m) = 0$ and, from *Lemma 2*, $c_{m-1,\mu} = c_{m,\mu}$ for $\mu = m + 1, \ldots, M$. Hence, $c_{0,1} = c_{0,2} = c_{1,2} = c_{0,3} = \ldots$, and generally $c_{i\mu} = 0$, $i < \mu$, $\mu = 1, \ldots, M$. But, then $\mathbf{C}$ and $\mathbf{C}^T$ are the Cholesky factors of $\mathbf{U}$ (for a definition see, for example, ref. 5).

COROLLARY 2: *Embedding of* $\mathbf{D}^2(\mathbf{U})$, *using the specific sequence* $d_{m-1,m}(m)$ *of generators, is identical to the Cholesky factorization of* $\mathbf{U}$.

Hence, the Cholesky factorization is a special case of embedding.

Numerical Performance of the Embedding Algorithm and Examples

As an example, we decompose the matrix $\mathbf{U} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which, when bordered by zeroes, becomes $\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$. The use of Eq. **37** to obtain the elements of $\mathbf{D}_2(1)$ yields

$$\mathbf{D}^2(\mathbf{U}) = \mathbf{D}^2(1) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -2 \\ 0 & -2 & 0 \end{pmatrix}. \quad [44]$$

Using d_{12} as the only possible (i.e., nonzero) generator, embedding (with Eq. **22**) yields $\mathbf{c}_1 = (c_{01}, c_{11}, c_{21}) = [-1/(2)^{1/2}, 0, -(2)^{1/2}]$ and $\varepsilon_1 = -1$; Eq. **23** leads to the matrix

$$\mathbf{D}^2(2) = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 0 \\ 1/2 & 0 & 0 \end{pmatrix} \quad [45]$$

and the vector $\mathbf{c}_2 = (c_{02}, c_{12}, c_{22}) = [0, (1/2)^{1/2}, (1/2)^{1/2}]$, $\varepsilon_2 = +1$. Translation of the origin P_0, whose coordinates are $(c_{01}, c_{02}) = [-(1/2)^{1/2}, 0]$, to the point (0, 0) is accomplished by a translation of $(1/2)^{1/2}$ applied to $\mathbf{c}_1$, and we obtain $\mathbf{c}_1 = [0, (1/2)^{1/2}, -(1/2)^{1/2}]$, $\mathbf{c}_2 = [0, (1/2)^{1/2}, (1/2)^{1/2}]$, and $\varepsilon_1 = -1$, $\varepsilon_2 = +1$. Thus, we have obtained a decomposition of **U** in the form of Eq. **41**. Hence, $p = q = 1$, and **U** has rank 2 and one negative and one positive eigenvalue. In this example, $\mathbf{c}_1 = [(1/2)^{1/2}, -(1/2)^{1/2}]$ and $\mathbf{c}_2 = [(1/2)^{1/2}, (1/2)^{1/2}]$, being the rows of **C**, are the eigenvectors of **U**.

We chose the above example to demonstrate an important fact. From *Theorem 5*, it follows that the sequence of generators $d_{m-1,m}(m)$ yields the factor **C** in triangular form, and the decomposition is equivalent to a Cholesky factorization of **U**. However, in this example, **U** cannot be factored by means of the Cholesky method because all diagonal elements are equal to zero. Nevertheless, the matrix **U** can be factored by the embedding of $\mathbf{D}^2(\mathbf{U})$, but **C** is not triangular.

Generally, if the diagonal elements of **U** are very small with respect to unity, then the Cholesky factorization cannot be guaranteed to be stable, as discussed by Gill *et al.* (6). For such matrices, the embedding of $\mathbf{D}^2(\mathbf{U})$ still produces a stable factorization. Consider the matrix (6)

$$\mathbf{V} = \begin{pmatrix} e & 1 \\ 1 & e \end{pmatrix} \text{ with } e \text{ close to zero.} \qquad [46]$$

Embedding of

$$\mathbf{D}^2(\mathbf{V}) = \begin{pmatrix} 0 & e & e \\ e & 0 & 2(e-1) \\ e & 2(e-1) & 0 \end{pmatrix} \qquad [47]$$

using d_{01} as the first generator, yields the vectors $(\varepsilon_1)^{1/2}\mathbf{c}_1 = (e^{1/2}, 1/e^{1/2})$ and $(\varepsilon_2)^{1/2}\mathbf{c}_2 = (0, [(e^2 - 1)/e]^{1/2})$. The factor **C** is triangular but the decomposition is very unstable. We obtain a stable factorization by using d_{12} as the generator in step 1; i.e., $\mathbf{c}_1 = (-[(e-1)/2]^{1/2}, [(e-1)/2]^{1/2}$, and $\mathbf{c}_2 = ([(e+1)/2]^{1/2}, [(e+1)/2]^{1/2})$.

Stability of the embedding algorithm is ensured if, in each step, m, the maximal element $d_{ik}^2(m) = \max\{|d_{jl}^2|, j, l = 0, 1, \ldots, N\}$, is used as the generator. Then

$$|c_{jm}| = |\overline{d}_{ij}(m)| = |d_{ij}^2(m) + d_{ik}^2(m) - d_{jk}^2(m)|/2|d_{ik}(m)| \leq \frac{3}{2}|d_{ik}(m)| \qquad [48]$$

and the magnitude of c_{jm} as well as the growth of $d_{ij}^2(m+1)$ is bounded in each step.

Embedding can be carried out very rapidly. The numbers of arithmetic operations required are $(N^3/6) + (N^2/2) + (N/3)$ multiplications, $(N^3/3) + N^2 + (2N/3)$ additions, and N square roots so that the computational complexity is $O(N^3)$, where N is the dimension of $\mathbf{D}^2$.

To evaluate the performance of the algorithm on a larger matrix, we calculated the embedding of the 58 × 58 distance matrix of the C^α atoms of the BPTI molecule. First, the matrix was generated using the coordinates obtained from the Protein Data Bank (7). Embedding of this distance matrix stopped correctly after three projections, yielding three coordinate vectors, with 58 components each. Recalculation of the distance matrix from these vectors showed no detectable error; i.e., $|d_{ij}^{orig} - d_{ij}^{calc}|$ was zero within the precision of the original coordinates (<0.001 Å). The execution time on a Prime 550 computer was 0.4 sec. For comparison, the embedding of an arbitrary 58 × 58 matrix $\mathbf{D}^2$ with full rank $M = 57$ (with its elements, between 0 and 1, generated randomly) took 7.1 sec for the 57 projections.¶

Conclusion

Recently, interest in the embedding problem has arisen from experimental techniques, especially in the field of two-dimensional NMR of biological macromolecules (8). Experimental techniques usually yield only partial information about the distances between atoms of a molecule, so that the distance matrices obtained are incomplete and the measured distances are known to lie in a certain range. A review of the recently developed techniques to solve the embedding of such matrices in euclidian three-dimensional space has been presented in ref. 9. Generally, one tries to derive a completely specified matrix that is consistent with the measured distances, by applying triangle and higher order inequalities [see also Braun *et al.* (10)]. The procedures used so far to obtain such a matrix do not guarantee that the matrix is embeddable in euclidian three-dimensional space. Two methods have been proposed to calculate coordinates from such matrices. MacKay (11) used the Cholesky factorization of the matrix **V** of Schoenberg's quadratic form (he and others use the term metric matrix for **V**). Since the Cholesky factorization may be unstable for indefinite matrices, he recommended the method only for positive **V** (11). Havel *et al.* (9) used the spectral decomposition of the matrix **V**. As stated by Schoenberg (3), the problem of finding a canonical representation of **V** is a problem of second degree because the coordinate vectors are not required to be orthogonal; hence, the embedding problem can be solved by more efficient techniques than by spectral decomposition. However, the spectral decomposition seems to be useful in reducing the dimensionality of **V** to obtain a three-dimensional structure.

The problem of calculating coordinates from an incompletely defined distance matrix involves the recovering or prediction of the missing or weakly defined elements of **D**. This problem, therefore, cannot be treated directly by the embedding procedure, but we hope that this problem will become more tractable through the analysis presented here.

We are indebted to E. O. Purisima for valuable discussions. This work was supported by the Stiftungs und Foerderungsgesellschaft der Paris Lodron Universitaet Salzburg, by the Max Kade Foundation, and by research grants from the National Institute of General Medical Sciences (GM-24893) and from the National Science Foundation (DMB84-01811). M.J.S. was a fellow of the Max Kade Foundation.

1. Menger, K. (1931) *Jahresber. Deutsch. Math.-Verein.* **40,** 201–219.
2. Blumenthal, L. M. (1970) *Theory and Applications of Distance Geometry* (Chelsea, New York), pp. 90–161.
3. Schoenberg, I. J. (1935) *Ann. Math.* **36,** 724–732.
4. Cartan, E. (1981) *Theory of Spinors* (Dover, New York), p. 4.
5. George, A. & Liu, J. W. H. (1981) *Computer Solution of Large Sparse Positive Definite Systems* (Prentice-Hall, Englewood Cliffs, NJ), pp. 15–25.
6. Gill, P. E., Murray, W. & Wright, M. H. (1981) *Practical Optimization* (Academic, London), pp. 36–37.
7. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
8. Wüthrich, K., Wider, G., Wagner, G. & Braun, W. (1982) *J. Mol. Biol.* **155,** 311–319.
9. Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983) *Bull. Math. Biol.* **45,** 665–720.
10. Braun, W., Bösch, C., Brown, L. R., Gō, N. & Wüthrich, K. (1981) *Biochim. Biophys. Acta* **667,** 377–396.
11. MacKay, A. L. (1983) in *Computing in Biological Sciences*, eds. Geisow, M. & Barret, A. (Elsevier Biomedical, Amsterdam), pp. 349–392.

¶Photocopies or microfiche of the computer program are available. See document no. NAPS 04264 of the ASIS National Auxiliary Publication Service, c/o Microfiche Publications, P.O. Box 3513, Grand Central Station, New York, NY 10017.