INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

# *Matrix Perturbation Theory in Distance Geometry*

Theodore G. Nikitopoulos

Juin 1999

# Matrix Perturbation Theory in Distance Geometry

Theodore G. Nikitopoulos *

Thème 2 — Génie logiciel
et calcul symbolique
Projet SAGA

Rapport de Maîtrise — Juin 1999

**Abstract:** The field of distance geometry has received lately an interest as it provides the mathematical basis for a geometric theory of molecular conformations. Quantum mechanics provides a detailed description of molecules in terms of atomic nuclei and electron distribution among them however, several approximated models have to be considered in practice. Therefore, in using a purely geometric model of molecular structure there are certainly limits but, there are also substantial advantages. In this work, the fundamental problem of distance geometry is studied, by using methods of matrix perturbation theory. In such approach we are interested to find a matrix perturbation such as every five-atom Cayley-Menger matrix is rank deficient. An implementation for the case of cycloalkanes is presented and further computational strategies are considered. One scope of this work is to propose an alternative strategy in interpreting the structural information obtained by Nuclear Magnetic Resonance methods. The framework of the theory suggest to promise well in identifying the geometry of small ligands, even though it is in a rather primitive state of development for the time being.

**Key-words:** Distance geometry, geometric algebra, molecular conformations, matrix perturbations, singular value decomposition, rank

* University of the Aegean, Department of Mathematics, 83200 Karlovassi, Samos, Greece, `s92035@math.aegean.gr`

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles, B.P. 93, 06902 Sophia Antipolis Cedex (France)
Téléphone : 04 92 38 77 77 - International : +33 4 92 38 77 77 — Fax : 04 92 38 77 65 - International : +33 4 92 38 77 65

# 1 Introduction

The mathematical framework of distance geometry belongs to the broad field of topology. The mappings—known as homeomorphisms—with which topology is principally concerned are biuniform and bicontinuous, but not, in general, distance-preserving. In distance geometry we deal with the specific subgroup of *congruences* or *isometries*, for which distance is invariant. The subject in which the extensive theory of this subgroup is developed is also known as *metric topology*. Its first systematic study was initiated by K. Menger in 1931 [Men31], when he formulated a new foundation of a geometry as the study of the invariants of figures under a selected group of transformations. A still worthwhile introduction to the subject, providing the historical roots of the field, is L. Blumenthal's text "Distance Geometry" [Blu70].

Identifying a molecular structure is a critical question in pharmaceutical drug design for human medicine with applications as well in other types of biological or chemical activities. Most drug discovery research is conducted with forethought and planning an argue even for scientists who screen huge numbers of compounds randomly selected. Chemists armed with information of a three-dimensional geometry, known from X-ray crystallography or Nuclear Magnetic Resonance (NMR), can think quilitatively about ligands that may bind to the active site of a macromolecule. Such knowledge is necessary in order to achieve an intuitive appreciation of the relationship between structure and drug activity [Kav97]. Structure-based methodologies as the above belong in the relatively new framework of computer-aided drug design.

However, in order to represent a molecular structure in the Euclidean space, someone has to interpret the structural information obtained by NMR methods. A molecule is determined by its lower and upper bounds of interatomic distances among its atoms—known as *distance constraints*. Chemists consider also *chirality constraints* which are the signs ($+1$, $-1$ or $0$) of the volumes of selected quadruples of atoms. The latter, determine the chirality of the quadruple or force it to be planar if the sign is zero. Usually, such information does not yield a single spatial structure, but rather an ensemble of different conformations that satisfy the constraints. It's therefore common to state that we calculate the *molecular conformations* of a specific molecule. Distance geometry provides a natural formulation of the above problem, which is also referred to as the *Fundamental Problem of Distance Geometry*:

> *Given the distance and chirality constraints which define (our state of knowledge of) a mobile molecule, find one or more molecular conformations which satisfy them, or else prove that no such conformations exist.*

Although X-ray crystallography is an alternative method of determining a molecular structure, in its case someone has to observe these molecules in an unnatural crystalline environment. Besides, it cannot be used in many cases because the molecule of our interest cannot be crystallized. Therefore, a large number of numerical methods exist in case of using NMR methods. Examples are the variable target function algorithm [BG85], the EMBED algorithm [HKC83, HHN97] and various simulated annealing algorithms [CKG86, SGK89] but, they all suffer from one or more shortcomings. However, the most important issue is that it happens in many cases to be unreliable. The above problem has been considered in the case of the structure determination of large molecules like proteins [Hav91]. It has been proposed that the areas in which drastic improvements may yet be possible is in the usual algorithmic steps of *bound smoothing* and of *embedding*. Our work is focused in the latter area but in a somewhat different direction.

Over the last few years a realization has been made that the underlying theory of distance geometry is closely related with a more general theory, known as *geometric algebra*. Although it has been proved to be useful in gaining insight into conformational problems, by finding closed form solutions of small molecules [Hav97], its scope is still limited as its expressions have to be reduced into complex systems of multivariate polynomials. Thus, in order to find isolated solutions of larger molecules there exists a tendency to use numerical methods, in locating global minima points, on artificially implemented *error functions*. Although, someone has a great deal of choices on the form of an error function—as to make it as smooth and well-behaved as possible—in our view, such a methodology disengages the geometric nature of the initial conformational problem. In this work, numerical methods are proposed in close relationship with the methods used in distance geometry. In a long term, any possible advancement in such a direction we believe to be conditioned by corresponding advancements in the theory of distance geometry.

## 2 Congruent embedding in Euclidean space

In this section, we review the mathematical problem of finding necessary and sufficient conditions in order for a molecule, defined by distance constraints, to be embeddable in the Euclidean space. This approach is based on the outline of L. Blumenthal's text [Blu70] and the formalism used by T. Havel in the text "Distance Geometry and Molecular Conformation" [CH88]. For the sake of brevity the reader is referred to the above citations for further discussion and detailed proofs.

Let $A$ be the set of atoms of a molecule, and $e\colon A \to \Re^n$ be a mapping[1] which associates with each $a \in A$ its Cartesian coordinates in a possible conformation of the molecule. By their definition these mappings are injective and we'll refer to them as *embeddings*. We denote by $E \subset (\Re^n)^A$ the set of all *admissible embeddings* which describe the possible conformations of the molecule of interest. For $e \in E$, we define the map $d_e\colon A \times A \to \Re$, which for $a, b \in A$ is given by:

$$d_e(a,b) = \|e(a) - e(b)\|$$

where $\|\mathbf{x}\| = \sqrt{\mathbf{x}\cdot\mathbf{x}}$ is the standard norm on the vector space $\Re^n$.

**Definition 2.1** *The map $d_e\colon A \times A \to \Re$ is called an $n$-dimensional Euclidean distance function.*

The following lemma derives some of the simplest properties of Euclidean distance functions.

**Lemma 2.1** *If $d_e$ is a Euclidean distance function and $D_e = d_e^2$, then we have:*
*1.* $\forall\, a, b \in A:\ D_e(a,b) = D_e(b,a)$
*2.* $\forall\, a, b \in A:\ D_e(a,b) = 0 \iff a = b$
*3.* $\forall\, a, b \in A:\ D_e(a,b) \geq 0$
*4.* $\forall\, a, b, c \in A:\ d_e(a,b) \leq d_e(a,c) + d_e(b,c)$

We now create terms for those real-valued functions $D$ defined on $A \times A$ which have just the above properties:

**Definition 2.2** *A* premetric space *$\mathbb{S}$ is a pair $(A, D)$ consisting of a set $A$ together with a function $D\colon A \times A \to \Re$ which satisfies conditions (1) and (2) of Lemma 2.1. If $D$ also satisfies condition (3), $\mathbb{S}$ is called a* semimetric space*, and if in addition the function $d = \sqrt{D}$ satisfies the triangle inequality (4), then $\mathbb{S}$ is called a* metric space*.*

We shall refer to the function $D$ itself as a *premetric*, whereas the function $d$ is called a *semimetric* or *metric*, as appropriate. The values $d(a,b)$ of $d$ are called *distances*, while the values $D(a,b)$ are called *squared distances*. Furthermore, if the size $|A|$ of $A$ is $m$ and we index and order the atoms as $\underline{A} = [\,a_1, \ldots, a_m\,]$, then the above spaces can be represented by means of a symmetric matrix:

[1] Although the dimension of the Euclidean space of our interest is always three, for the time being we'll consider the general case.

$$\underline{\mathbf{M}}_D(\underline{A}) = \begin{bmatrix} 0 & D(a_1,a_2) & \dots & D(a_1,a_m) \\ D(a_1,a_2) & 0 & \dots & D(a_2,a_m) \\ \vdots & \vdots & \ddots & \vdots \\ D(a_1,a_m) & D(a_2,a_m) & \dots & 0 \end{bmatrix}$$

Thus a finite premetric space can be identified with the set of all matrices obtained from $\underline{\mathbf{M}}_D(\underline{A})$ by row/column permutation, and similarly semimetric and metric spaces can be regarded as being essentially a certain type of matrix.

**Definition 2.3** *A premetric space $\mathbb{S} = (A, D)$ is* Euclidean *if there exists an embedding $e : A \rightarrow \Re^n$ such that $D_e = d_e^2$, the square of the corresponding Euclidean distance function. We may also describe this by saying that $\mathbb{S}$ is* embeddable *in an $n$-dimensional Euclidean space.*

An Euclidean premetric space by its definition is a metric space. However, even though they obey the triangle inequality there exist metric spaces $(A, D)$ which are not Euclidean. We'll seek an analogous algebraic representation which is applicable to any premetric space.

**Definition 2.4** *A premetric space $\mathbb{S} = (A, D)$ is said to be* embeddable *in a bilinear affine space $\mathbb{A}_B$ with quadratic form $Q_B$ if there exists an embedding $e : A \rightarrow \mathbb{A}_B$ such that $D(a, b) = Q_B(e(a) - e(b))$ for all atoms $a, b \in A$. The embedding $e$ itself is called an* isometric embedding *of $\mathbb{S}$ in $\mathbb{A}_B$. If such an embedding exists, the dimension $dim(\mathbb{S})$ of $\mathbb{S}$ is then the minimum dimension of a bilinear affine space in which it is embeddable.*

**Theorem 2.1** *Every premetric space $\mathbb{S} = (A, D)$ can be embedded in a bilinear affine space of dimension $|A| - 1$ such that the image of $A$ spans the space. This embedding is unique up to isometry.*

**Proof:** We'll construct this space using barycentric coordinates. We enlarge the set $A$ to a set $\tilde{\Re}[A]$ consisting of all *formal sums* $\sum \kappa_i a_i$ of the elements in $A$ with $\sum \kappa_i = 1$. We then define a vector space $\hat{\Re}[A]$ whose elements are all formal sums such that $\sum \lambda_i = 0$. Vector addition and scalar multiplication are defined as:

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i a_i + \sum_{i=1}^{m} \lambda'_i a_i &= \sum_{i=1}^{m} (\lambda_i + \lambda'_i) a_i \\ \mu \cdot \sum_{i=1}^{m} \lambda_i a_i &= \sum_{i=1}^{m} (\mu \lambda_i) a_i \end{aligned}$$

where $m = |A|$. Because of the constraint $\sum \lambda_i = 0$, the dimension of $\hat{\Re}[A]$ is equal to $|A| - 1$. Furthermore, if we define a map from $\tilde{\Re}[A] \times \hat{\Re}[A]$ to $\tilde{\Re}[A]$ according to the rule:

$$\left( \sum_{i=1}^{m} \kappa_i a_i \, , \, \sum_{i=1}^{m} \lambda_i a_i \right) \mapsto \sum_{i=1}^{m} (\kappa_i + \lambda_i) a_i$$

it is verified that $\mathbb{A}[A] = (\tilde{\Re}[A], \hat{\Re}[A])$, together with this map, fulfills all the axioms for an affine space. The translation which maps $\sum \kappa_i a_i$ to $\sum \kappa_i' a_i$ is obtained by subtracting corresponding coefficients to obtain $\sum \lambda_i a_i$ with $\sum \lambda_i = \sum (\kappa_i - \kappa_i') = 0$. We call $\mathbb{A}[A]$ the *amalgamation space* of A. We now define a bilinear form on $\hat{\Re}[A]$ according to the equation:

$$B_D \left( \sum \lambda_i a_i \, , \, \sum \lambda_j' a_j \right) = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} D(a_i, a_j) \lambda_i \lambda_j'$$

The corresponding quadratic form $Q_D \colon \hat{\Re}[A] \to \Re$ is known as *Schoenberg's quadratic form*, and the bilinear vector space $(\hat{\Re}[A], B_D)$ will be denoted by $\hat{\Re}_D[A]$. If we define our embedding $e \colon A \to \tilde{\Re}[A]$ as $e(a_j) = \sum \delta_{ij} a_i$, where $\delta_{ij}$ is the Kronecker delta, then we have:

$$Q_D(e(a_i) - e(a_j)) = D(a_i, a_j)$$

Thus, $e$ is the desired isometric embedding of $A$ in the bilinear affine space $\mathbb{A}_D[A] = (\tilde{\Re}[A], \hat{\Re}[A], B_D)$. We call $e$ the *canonical embedding* in $\mathbb{A}_D[A]$. It is easily verified that this embedding is unique up to isomorphism. □

**Corollary 2.1** *A premetric space $\mathbb{S} = (A, D)$ has dimension $n$ if and only if Schoenberg's quadratic form $Q_D$ is of rank $n$ on $\hat{\Re}[A]$. In addition, a premetric space is Euclidean if and only if the form is positive semidefinite.*

The above Corollary provides us with a necessary and sufficient condition for a premetric space to be Euclidean of dimension $n$. However, we need an insight into how the signature of the form depends on the values of the individual squared distances $D(a, b)$.

**Definition 2.5** *The* Cayley-Menger determinant *of a premetric space* $(A, D)$*, with* $|A| = m$*, is* $D(\underline{A}) = 2(-1/2)^m \det(\underline{\tilde{\mathbf{M}}}_D(\underline{A}))$*, where* $\underline{\tilde{\mathbf{M}}}_D(\underline{A})$ *is the bordered matrix:*

$$\underline{\tilde{\mathbf{M}}}_D(\underline{A}) = \begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & D(a_1, a_2) & \dots & D(a_1, a_m) \\ 1 & D(a_1, a_2) & 0 & \dots & D(a_2, a_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & D(a_1, a_m) & D(a_2, a_m) & \dots & 0 \end{bmatrix} \quad (1)$$

Obviously this determinant is independent of the order in which we number our atoms, thus we can refer to *the* determinant of the space without ambiguity. When the premetric is clear, we may also refer to $D(\underline{A}) = D(A)$ as the Cayley-Menger determinant of the atoms $A$ themselves.

**Lemma 2.2** *Let* $(A, D)$ *be a premetric space,* $D(A)$ *be its Cayley-Menger determinant,* $\underline{A} = [\, a_1, \dots, a_m \,]$ *be an ordering of* $A$*, and* $[\, \mathbf{v}_j = \mathbf{a}_j - \mathbf{a}_1 \,|\, j = 2, \dots, m \,]$ *be the basis of the amalgamation space* $\mathbb{A}_D$ *obtained by choosing* $\mathbf{a}_1$ *as the origin. Then*

$$D(A) \;=\; D(a_1, \dots, a_m) \;=\; B_D(\mathbf{v}_2, \dots, \mathbf{v}_m),$$

*where* $B_D(\mathbf{v}_2, \dots, \mathbf{v}_m)$ *is the determinant of the metric matrix* $\underline{\mathbf{M}}_B(\mathbf{v}_2, \dots, \mathbf{v}_m)$*:*

$$\underline{\mathbf{M}}_B(\mathbf{v}_2, \dots, \mathbf{v}_m) = \begin{bmatrix} B(\mathbf{v}_2, \mathbf{v}_2) & B(\mathbf{v}_2, \mathbf{v}_3) & \dots & B(\mathbf{v}_2, \mathbf{v}_m) \\ B(\mathbf{v}_2, \mathbf{v}_3) & B(\mathbf{v}_3, \mathbf{v}_3) & \dots & B(\mathbf{v}_3, \mathbf{v}_m) \\ \vdots & \vdots & \ddots & \vdots \\ B(\mathbf{v}_2, \mathbf{v}_m) & B(\mathbf{v}_3, \mathbf{v}_m) & \dots & B(\mathbf{v}_m, \mathbf{v}_m) \end{bmatrix} \quad (2)$$

**Theorem 2.2** *A premetric space* $\mathbb{S} = (A, D)$ *is Euclidean of dimension* $dim(\mathbb{S}) = n > 0$ *if and only if for all* $B \subseteq A$ *with* $n + 1 \geq |B| > 1$ *we have* $D(B) \geq 0$*, while* $D(B) = 0$ *for all* $B \subseteq A$ *with* $|B| > n + 1$*, and* $n$ *is the minimum positive integer with this property.*

**Proof:** For $B = \{b_1, \dots, b_k\} \subseteq A$, let $\underline{\mathbf{U}} = [\, \mathbf{b}_j - \mathbf{b}_1 \,|\, j = 2, \dots, m \,]$, and recall that a bilinear form $B : \mathbb{V} \times \mathbb{V} \to \Re$ is positive (semi)definite if $\forall\, \mathbf{u} \in \mathbb{V}$, $B(\mathbf{u}, \mathbf{u}) \geq 0$ whereas $B(\mathbf{u}, \mathbf{u}) = 0$ implies $\mathbf{u} = 0$. Since $D(B) = B(\mathbf{U}) = \det(\underline{\mathbf{M}}_B(\underline{\mathbf{U}}))$ by Lemma 2.2 and any subspace of a positive semidefinite bilinear vector space is obviously positive semidefinite, it suffices to show that the discriminant of a positive (semi)definite bilinear vector space is +1 (or 0). This is easily seen by reducing the metric matrix $\underline{\mathbf{M}}_B(\underline{\mathbf{U}})$ to canonical form, i.e.

$$B(\underline{\mathbf{U}}) \;=\; \det(\underline{\mathbf{X}}^T \mathbf{\Omega}\, \underline{\mathbf{X}}) \;=\; \det{}^2\,(\underline{\mathbf{X}}) \cdot \prod_{i=2}^{k} \omega_{ii} \;\geq\; 0$$

since $\omega_{ii} \in \{0, +1\}$ for $2 \leq i \leq k$.

The converse follows from Corollary 2.1 and a well-known relation between positive (semi)definite quadratic forms and symmetric matrices. □

In Theorem 2.2 it's stated that the nonnegativity of any symmetric two-, three-, and four-atom Cayley-Menger determinants, together with the vanishing of all higher determinants is sufficient for a premetric space to be three-dimensional Euclidean. However, a weakened condition also exists in which—assuming that none of the distances are zero—it's sufficient to vanish all five-atom Cayley-Menger determinants. The latter fact implies that any five-atom Cayley-Menger matrix of type (1) should have at most rank five or in other terms that it should be rank-deficient.

## 3 Singular value decomposition

The theme of diagonalizing a matrix by expressing it in terms of a new basis is of a major importance in the theory of matrices. The history of singular value decomposition is dated back in nineteenth-century [Ste93] when differential geometers and algebraists wanted to know how to determine whether two real bilinear forms:

$$\varphi_A(x,y) = \sum_{i,j=1}^{n} a_{ij} x_i y_j \quad \varphi_B(x,y) = \sum_{i,j=1}^{n} b_{ij} x_i y_j$$
$$A = [a_{ij}],\ B = [b_{ij}] \in \Re^{n\times n},\ x = [x_i],\ y = [y_i] \in \Re^n$$

were equivalent under independent real orthogonal substitutions, that is, whether there are real orthogonal $Q_1,\ Q_2 \in \Re^{n\times n}$ such that $\varphi_A(x,y) = \varphi_B(Q_1 x, Q_2\, y)$ for all $x, y \in \Re^n$. One can approach this problem by finding a canonical form to which any such bilinear form can be reduced by orthogonal substitutions.

In 1873, the Italian differential geometer E. Beltrami discovered that for each $A \in \Re^{n\times n}$ there are always real orthogonal $Q_1,\ Q_2 \in \Re^{n\times n}$ such that:

$$Q_1^T A Q_2 = \Sigma = \operatorname{diag}(\sigma_1(A), \ldots, \sigma_n(A)) \tag{3}$$

is a nonnegative diagonal matrix, where $\sigma_1(A)^2 \geq \cdots \geq \sigma_n(A)^2$ are the eigenvalues of both $AA^T$ and $A^TA$ while the columns of $Q_1$ and $Q_2$ are the eigenvectors of $AA^T$ and $A^TA$ respectively. This canonical form is uniquely determined and it's called the *singular value decomposition* of a real square matrix while, the *singular values* of $A$ are the numbers $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$. The diagonal bilinear form:

$$\varphi_{Q_1^TAQ_2}(\xi,\eta) = \sigma_1(A)\xi_1\eta_1 + \cdots + \sigma_n(A)\xi_n\eta_n$$

gives a convenient canonical form to which any real bilinear form $\varphi_A(x,y)$ can be reduced by independent orthogonal substitutions. In the following we review some classic results of the singular value decomposition of any real square matrix [GL96].

**Theorem 3.1** *Let $A \in \Re^{n\times n}$ be given, and let $Q_1, \Sigma, Q_2$ be the corresponding matrices of the singular value decomposition of $A$. Then we have that $A = Q_1\Sigma Q_2^T$ and moreover the rank of $A$ is specified by the number of its nonzero singular values.*

The above implies that the matrix $A$ can be represented, as a sum of $r$ rank-one matrices, by a form:

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^T \qquad r = \text{rank(A)}$$

where $u_k$ and $v_k$ are the $k$-th columns of $Q_1$ and $Q_2$ respectively. Although a square matrix $A$ can be expressed as a sum of rank-one matrices in many ways, the above formula suggest an interesting decomposition:

**Theorem 3.2** *If we define the $n$-th partial sum of a matrix $A \in \Re^{n\times n}$ as:*

$$A_n = \sum_{j=1}^{n} \sigma_j u_j v_j^T \qquad 0 \leq n \leq r \tag{4}$$

*and if $n = r$ we define $\sigma_{r+1} = 0$ then, such a matrix satisfies a best approximation of $A$, by matrices of lower rank, in respect of the 2-norm or the Frobenius norm, i.e.*

$$\begin{aligned}
\|A - A_n\|_2 &= \inf_{\substack{B\in\Re^{n\times n}\\ rank(B)\leq n}} \|A - B\|_2 = \sigma_{n+1} \\
\|A - A_n\|_F &= \inf_{\substack{B\in\Re^{n\times n}\\ rank(B)\leq n}} \|A - B\|_F = \sqrt{\sigma_{n+1}^2 + \cdots + \sigma_r^2}
\end{aligned}$$

An also important question in the theory of matrices is concerned with the least squares solution of overdetermined systems of equations, i.e, the minimization of $\|Ax - b\|_2$ where $A \in \Re^{n\times n}$ is rank-deficient and $b \in \Re^n$. The most reliable solution procedures for this problem involve the reduction of $A$ to various canonical forms via orthogonal transformations.

**Definition 3.1** *Let $A \in \Re^{n\times n}$ and suppose that its singular value decomposition is given by $A = Q_1 \Sigma Q_2^T$, then the matrix $A^+ = Q_2 \Sigma^+ Q_1^T$ where*

$$\Sigma^+ = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_\mathrm{r}}, 0, \ldots, 0\right) \in \Re^{\mathrm{n}\times\mathrm{n}} \qquad \mathrm{r} = \mathrm{rank(A)}$$

*is referred as the* pseudo-inverse *of $A$. Typically, $A^+$ is defined to be the unique matrix $X \in \Re^{n\times n}$ that satisfies the four* Moore-Penrose conditions*:*

$$\begin{array}{llll} 1. & AXA = A & 3. & (AX)^T = AX \\ 2. & XAX = X & 4. & (XA)^T = XA \end{array}$$

**Theorem 3.3** *Suppose $A = Q_1 \Sigma Q_2^T$ is the singular value decomposition of the matrix $A \in \Re^{n\times n}$. Then the least squares solution $x_{LS} = A^+ b$ minimizes the $\|Ax - b\|_2$ and has the smallest 2-norm of all minimizers.*

The reader will probably notice a close relationship between the singular value decomposition of a matrix $A$ and its rank. We are interested on the ability of it to handle rank-deficiency in the presence of a roundoff error.

**Theorem 3.4** *Let $A \in \Re^{n\times n}$ be given and let $\Delta A \in \Re^{n\times n}$ such that $\|\Delta A\|_2 \leq \epsilon \|A\|_2$, where $\epsilon > 0$. We denote the matrix $\hat{A}$ as $\hat{A} = A + \Delta A$, then we have that $|\sigma_k(A) - \sigma_k(\hat{A})| \leq \epsilon\, \sigma_1(A)$ for any $k \in \{1, \ldots, n\}$.*

The above suggest that if A has rank $r$ then we can expect $n - r$ of the computed *numerical singular values* to be small. In such a case $\epsilon$ should correspond to a machine's precision. Nevertheless, it give us also an insight into a deeper property. In case singular values are not viewed as isolated points, but as values of a function defined by $\sigma_k(A - R)$ as $R$ varies over a field $\mathcal{R}$ then this function is continuous. Therefore, a form of differentiation of such a function could be possible.

**Theorem 3.5** *Let $A \in \Re^{n \times n}$ be given and let $R \in \Re^{n \times n}$ such as the $n$-th singular value of $A - R$ is distinct and nonzero. Then the function defined by:*

$$f(a) \stackrel{\Delta}{=} \sigma_n(A - aR) \qquad a \in \Re$$

*is differentiable with respect to $a$ and its derivative is given by $f'(a) = -u_n R v_n$ where $u_n$ and $v_n$ correspond to the singular vector pair associated with $\sigma_n(A - aR)$.*

In the above case, a Newton iteration is obtained by the following form:

$$\begin{aligned} a_{k+1} &= a_k + [u_k^T(A - a_k R)v_k]/(u_k R v_k) \\ &= (u_k^T A v_k)/(u_k^T R v_k) \end{aligned} \tag{5}$$

where $u_k$ and $v_k$ are the singular vectors associated with $\sigma_n(A - a_k R)$. Although we've assumed that the dimension of $\mathcal{R}$ is one, a generalization also exists in which $\mathcal{R}$ can be multidimensional. Given any particular basis of the form $\{R_i\}$, $i = 1, \ldots, s$ over this field a matrix is expressed as $R = \sum_{i=1}^{s} R_i \xi_i$ and a Newton iteration such that of type (5) can be obtained by locating $\Delta R_k \in \mathcal{R}$ that should satisfy the overdetermined system:

$$u_k^T(\Delta R_k)v_k = u_k^T(A - R_k)v_k$$

and by setting $R_{k+1} = R_k + a_k \Delta R_k$, where $a_k$ is given by:

$$a_k = (u_k^T(A - R_k)v_k)/(u_k^T \Delta R_k v_k)$$

and $u_k$, $v_k$ are the singular vectors associated with $\sigma_n(A - R_k)$.

In other terms, a numerical approach in finding a rank-reducing matrix perturbation or a local minimum for a most nearly rank-reducing perturbation of any initial matrix $A$ of full-rank exist [WD95]. A worthwhile introduction in this broad field is G. Stewart's and Ji-guang Sun's text "Matrix Perturbation Theory" [SS90].

# 4 Algorithmic approaches

Our initial interest is to implement an algorithm where given an arbitrary premetric space we construct a new premetric space, in which *every* five-atom Cayley-Menger matrix of the molecule's is rank-deficient or most nearly rank-deficient. In this section we discuss a possible approach.

**Step 1.** Let $A$ be a set of the atoms of a molecule and let $|A| = m$ is the number of its atoms. We construct a premetric space $\mathbb{S} = (A, D)$ such as $D$ satisfies the squared values of the distance constraints.

**Step 2.** We suppose that $A$ is ordered, and we denote a set $\mathcal{C}(A)$ such as:

$$\mathcal{C}(A) = \{B \,:\, B \subset A\,,\ |B| = 5\}$$

Obviously for every $B \in \mathcal{C}(A)$ there exists a set $D_B \subset D$ such as the Cayley-Menger matrix of $B$ is defined. We construct a field $\mathcal{R}$ such as its base is defined by matrices of the form $\{R_i\}$, $i = 1, \ldots, m^2/2 - m$ having zero-entries except of ones in each $(k, l)$ position and its symmetric one associated in any different pair of atoms $a_k$ and $a_l$.

**Step 3.** We find the most bad behaved five-atom Cayley-Menger matrix on being rank-deficient. This is suggested by its sixth singular-value, which is the largest of any other's five-atom Cayley-Menger matrix. We denote it by:

$$\Sigma_{max} = \max\{\,\sigma_6(\underline{\tilde{\mathbf{M}}}(B)) \;:\; B \in \mathcal{C}(A)\}$$

This value implies actually a starting point into a better conditioned system.

**Step 4.** Let $B \in \mathcal{C}(A)$, we find a perturbation of its Cayley-Menger matrix such as its sixth singular-value is reduced and satisfies the distance constraints. The perturbed matrix is provided by a *one-step* multidimensional Newton like iteration, over the field $\mathcal{R}$. We construct a temporary premetric space $\mathbb{S}' = (A, D')$ where $D'$ is suggested by the perturbed matrix.

**Step 5.** For any $B \in \mathcal{C}(A)$ we check if the inequality $\sigma_6(\underline{\tilde{\mathbf{M}}}(B)) < \Sigma_{max}$ is satisfied in the premetric space $\mathbb{S}'$ otherwise, we go back to Step 4 untill the inequality holds for a new premetric or untill any $B \in \mathcal{C}(A)$ has been perturbed. If the latter case holds then we stop, otherwise we go back to Step 3.

A couple of remarks should be made in following such a direction. Although Theorem 3.5 can be also applied in other singular values apart the last one, certain problem structures cause an algorithm to disbehave. It's a common case when someone tries to reduce the sixth singular value[1] of the Cayley-Menger matrix $\underline{\tilde{\mathbf{M}}}(A)$ some

[1] Actually another condition also exists in which a premetric's space Cayley-Menger matrix is sufficient to be at most of rank five.

smaller singular value can actually become greater. Such a behaviour is expected as the nature of a derivative is relied on local properties. Therefore, in order to handle any various appearing 'small' oscillations an approach based on an heuristic implementation has to be assumed. Even though it's possible to provide solutions on small molecules such a direction is not necessary sufficient to handle larger problems without any apropriate modifications.

An also important issue is that the mathematical framework of the theory doesn't presuppose that a perturbed matrix should satisfy any predefined distance constraints. As a result, although it's relatively easy to construct a matrix of whatever rank—actually Theorem 3.2 suggests a way—for the specific problem of distance geometry such solutions are not of any significant importance. For the time being it's an easier case to satisfy the distance constraints on perturbed matrices of small size, i.e. the case of any five-atom Cayley-Menger matrix. An intuition of the range between any lower and upper distance bound of a 83-atom molecule is given in the Appendix.

While considering every five-atom Cayley-Menger matrix is also computationally intensive on big molecules, till now it provides a good insight on the specific nature of the related questions. Such an insight we hope to suggest also further directions on attaching the conformational problem as a whole.

## 5 A case study of Cycloalkanes

The algorithms discussed in previous section make no assumption about the geometry of molecular chains. However we applied them to the specific problem of ring closure as it's well-studied by analytical [Hav97] or by a combination of analytical and numerical methods [EM99]. Therefore, we had a broad reference of literature in comparing the results of our approach.

### 5.1 Cyclohexane conformations

Cyclohexane has an infinite geometrically possible confomation due to its structural symmetry. Actually it can assume a rigid chair conformation as well as a flexible 'circle' of conformers related to the boat form. In its case, energy minima are typically found among the conformations obtained by keeping bond lengths and bond angles fixed and allowing only the torsion dihedral angles to vary about single covalent bonds. The above are represented by considering distances for the atoms involved and they are referred as the 1-2, 1-3 and 1-4 distances respectively.

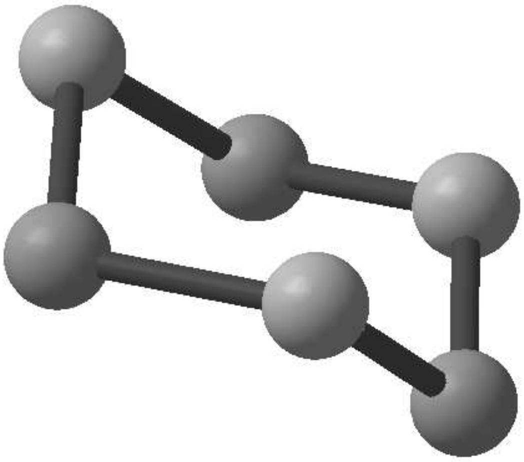

Figure 1: One conformation of a cyclic 6-atom molecule.

In the following example, we slightly decrease the length of the first bond. Since the symmetry no longer holds, we can obtain a finite set of solutions. Using the procedure described in Section 4, we've found four solutions. The first two correspond to the chair conformation and the other two correspond to those of the twisted boat conformation. The Cayley-Menger matrix of cyclohexane while considering the torsion dihedral angles as unknowns is given by:

$$\underline{\tilde{\mathbf{M}}}(A) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & c_{1,2} & c_{1,3} & \mathbf{u}_1 & c_{1,5} & c_{1,6} \\ 1 & c_{1,2} & 0 & c_{2,3} & c_{2,4} & \mathbf{u}_2 & c_{2,6} \\ 1 & c_{1,3} & c_{2,3} & 0 & c_{3,4} & c_{3,5} & \mathbf{u}_3 \\ 1 & \mathbf{u}_1 & c_{2,4} & c_{3,4} & 0 & c_{4,5} & c_{4,6} \\ 1 & c_{1,5} & \mathbf{u}_2 & c_{3,5} & c_{4,5} & 0 & c_{5,6} \\ 1 & c_{1,6} & c_{2,6} & \mathbf{u}_3 & c_{4,6} & c_{5,6} & 0 \end{bmatrix}$$

where all the coefficients $c_{i,j} = d(a_i, a_j)^2$ are known from the input parameters $L_i$, $\phi_i$ listed in Table 1 and there are only three unknowns $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$. Once the unknows are determined, we can recover the geometry of the molecule up to global translations and rotations as illustrated in Table 2.

| Number $i$ | Link Length $a_i$ | Offset Distance $L_i$ | Twist Angle $\phi_i$ |
|---|---|---|---|
| 1 | 0.0 | 1.525 | 69.6 |
| 2 | 0.0 | 1.526 | 69.6 |
| 3 | 0.0 | 1.526 | 69.6 |
| 4 | 0.0 | 1.526 | 69.6 |
| 5 | 0.0 | 1.526 | 69.6 |
| 6 | 0.0 | 1.526 | 69.6 |

Table 1: Input parameters of the cyclohexane molecule.

| $i$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ |
|---|---|---|---|---|---|---|
| 1 | -57.69 | 57.67 | -57.62 | 57.60 | -57.62 | 57.67 |
| 2 | 57.69 | -57.67 | 57.62 | -57.60 | 57.62 | -57.67 |
| 3 | 67.78 | -32.04 | -31.98 | 67.70 | -31.98 | -32.04 |
| 4 | -67.78 | 32.04 | 31.98 | -67.70 | 31.98 | 32.04 |

Table 2: The dihedral angles of cyclohexane after recovering its geometry

## 5.2 Cyclooctane conformations

In the case of cyclooctane there are eight rotable bonds and there exist twelve unknowns. Again once the unknows are determined, we can recover the geometry of the molecule up to global translations and rotations. The Cayley-Menger matrix of cyclooctane is given by:

$$\underline{\tilde{\mathbf{M}}}(A) = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & c_{1,2} & c_{1,3} & \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_3 & c_{1,7} & c_{1,8} \\ 1 & c_{1,2} & 0 & c_{2,3} & c_{2,4} & \mathbf{u}_4 & \mathbf{u}_5 & \mathbf{u}_6 & c_{2,8} \\ 1 & c_{1,3} & c_{2,3} & 0 & c_{3,4} & c_{3,5} & \mathbf{u}_7 & \mathbf{u}_8 & \mathbf{u}_9 \\ 1 & \mathbf{u}_1 & c_{2,4} & c_{3,4} & 0 & c_{4,5} & c_{4,6} & \mathbf{u}_{10} & \mathbf{u}_{11} \\ 1 & \mathbf{u}_2 & \mathbf{u}_4 & c_{3,5} & c_{4,5} & 0 & c_{5,6} & c_{5,7} & \mathbf{u}_{12} \\ 1 & \mathbf{u}_3 & \mathbf{u}_5 & \mathbf{u}_7 & c_{4,6} & c_{5,6} & 0 & c_{6,7} & c_{6,8} \\ 1 & c_{1,7} & \mathbf{u}_6 & \mathbf{u}_8 & \mathbf{u}_{10} & c_{5,7} & c_{6,7} & 0 & c_{7,8} \\ 1 & c_{1,8} & c_{2,8} & \mathbf{u}_9 & \mathbf{u}_{11} & \mathbf{u}_{12} & c_{6,8} & c_{7,8} & 0 \end{bmatrix}$$

Even though our procedure cannot find solutions for the time being, we can recover the geometry of the molecule by using heuristic techniques. A couple of states of all the five-atom Cayley-Menger matrices after using the procedure of Section 4 is illustrated in the Appendix.

## 6 Conclusions

The fundamental problem of distance geometry appears to be of a major importance in pharmaceutical industry and furthermore of a mathematical challenge by itself. However the reader should be noted that this work is actually in a rather primitive state of development for the time being. Certain issues have to be considered on advantages and disadvantages of each alternative approach. An important remark for the time is that actually there exist direct numerical methods on this particular problem which can also consider its characteristic geometric nature.

## Acknowledgment

I would like to thank Ioannis Z. Emiris, my thesis advisor, for many insightful suggestions. Most of this work was conducted as part of a project of SAGA group in INRIA Sophia-Antipolis. My gratitude for their hospitality and their support.

## References

[BG85] W. Braun and N. Gō. Calculation of protein conformations by proton-proton distance constraints. An efficient new algorithm. *J. Molec. Biol.*, 186:611–626, 1985.

[Blu70] L. M. Blumenthal. *Theory and Applications of Distance Geometry.* Chelsea Publishing Company, Bronx, NY, 1970.

[CH88] G. M. Crippen and T. F. Havel. *Distance Geometry and Molecular Conformation.* Chemometrics series, Vol. 15. J. Wiley and sons, 1988.

[CKG86] G. M. Clore, A. T. Karplus, and A. M. Gronenborn. Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination: A model study of crambin. *J. Molec. Biol.*, 191:523–551, 1986.

[EM99] I. Z. Emiris and B. Mourrain. Computer algebra methods for studying and computing molecular conformations. *Algorithmica*, pages 1–32, to appear.

[GL96] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 3rd. edition, 1996.

[Hav91] T. F. Havel. An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Molec. Biol.*, 56:43–78, 1991.

[Hav97] T. F. Havel. Computational synthetic geometry with Clifford algebra. In *Automated deduction in geometry*, volume 1360 of *LNCS*, pages 102–114. Springer, 1997.

[HHN97] T. F. Havel, S. Hyberts, and I. Najfeld. Recent advances in molecular distance geometry. In *Bioinformatics*, volume 1278 of *LNCS*, pages 62–71. Springer, 1997.

[HKC83] T. F. Havel, I. D. Kuntz, and G. M. Crippen. Theory and practice of distance geometry. *Bull. Math. Biol.*, 45:665–720, 1983.

[Kav97] L. E. Kavraki. Geometry and the discovery of new ligands. In *Algorithms for Robotic Motion and Manipulation*, pages 435–448. A. K. Peters, 1997.

[Men31] K. Menger. New foundation of Euclidean geometry. *American Journal of Mathematics*, 53:721–745, 1931.

[SGK89] R. M. Scheek, W. F. Gunsteren, and R. Kaptein. Molecular dynamics simulation techniques for determination of molecular structures from nuclear magnetic resonance data. *Meth. Enzymol.*, 177:204–218, 1989.

[SS90] G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, Inc., San Diego, 1990.

[Ste93] G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35:551–566, 1993.

[WD95] M. A. Wicks and R. A. Decarlo. Computing most nearly rank-reducing structured matrix perturbations. *SIAM J. Matrix Anal. and Appl.*, 16:123–137, 1995.

Unité de recherche INRIA Sophia Antipolis
2004, route des Lucioles - B.P. 93 - 06902 Sophia Antipolis Cedex (France)

Unité de recherche INRIA Lorraine : Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 Villers lès Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38330 Montbonnot St Martin (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 Le Chesnay Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, B.P. 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr