Journal of Mathematical Chemistry 21 (1997) 223–260

Embedding with a rigid substructure

Igor Najfeld and Timothy F. Havel
BCMP, Harvard Medical School, Boston, MA 02115, USA

Received 11 March 1997

This paper presents a new distance geometry algorithm for calculating atomic coordinates from estimates of the interatomic distances, which maintains the positions of the atoms in a known rigid substructure. Given an $M \times 3$ matrix of coordinates $\mathbf{X}$ for the rigid substructure, this problem consists of finding the $N \times 3$ matrix $\mathbf{Y}$ which yields the global minimum of the so-called *STRAIN*, i.e.,

$$\min_{\mathbf{Y}} \left\| \begin{bmatrix} \mathbf{X}\mathbf{X}^\top & \mathbf{X}\mathbf{Y}^\top \\ \mathbf{Y}\mathbf{X}^\top & \mathbf{Y}\mathbf{Y}^\top \end{bmatrix} - \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C} \end{bmatrix} \right\|_{\mathrm{F}}^2,$$

where $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$, and $\mathbf{B}$, $\mathbf{C}$ are matrices of inner products calculated from the estimated distances.

The vanishing of the gradient of the *STRAIN* is shown to be equivalent to a system of only six nonlinear equations in six unknowns for the inertial tensor associated with the solution $\mathbf{Y}$. The entire solution space is characterized in terms of the geometry of the intersection curves between the unit sphere and certain variable ellipsoids. Upon deriving tight bilateral bounds on the moments of inertia of any possible solution, we construct a search procedure that reliably locates the global minimum. The effectiveness of this method is demonstrated on realistic simulated and chemical test problems.

1. Introduction

Distance geometry is widely used in the generation of three-dimensional atomic coordinates from chemical diagrams [7], in the determination of conformation from NMR data [17], in the modeling of protein structure by homology [16], and in docking ligands in their receptor's binding site [4]. More generally, distance geometry is a formalism within which a large variety of conformational problems can be stated in purely geometric terms [9]. Solving these problems generally involves finding three-dimensional coordinates that satisfy given lower and upper bounds on the interatomic distances. In contrast to relatively complex models of the intramolecular potential energy, this simple geometric problem formulation makes it possible to use rigorous mathematical techniques to assist the search for globally optimal solutions.

The most important of these mathematical techniques is the *EMBED* algorithm, which rapidly finds coordinates that closely fit the distance bounds [8]. As explained in the next section, these coordinates are the global minimum of a certain matrix optimiza-

© J.C. Baltzer AG, Science Publishers

tion problem, which can be reliably located by eigenvalue methods. The availability of such good starting coordinates greatly improves the efficiency and convergence properties of their subsequent nonlinear optimization versus an *error function*, which measures the violations of the distance bounds [15]. One shortcoming of the *EMBED* algorithm is that there is no good way to weight the distances according to their precision, or exactly realize rigid substructures even when they are known in advance. This shortcoming is particularly serious in problems with large rigid substructures, as in the docking of a ligand to a receptor protein of known structure, or in building the loops and sidechains of a protein onto a backbone obtained via homology modeling.

In this paper we formulate a new minimization criterion which fixes the positions of the atoms in a known rigid substructure. It may be regarded as a block matrix extension of the *EMBED* algorithm. We derive the necessary conditions for an optimal solution in the form of a nonlinear matrix equation. By considering the inertial tensor associated with a solution matrix, it is possible to transform this matrix equation into a system of six nonlinear equations among the three eigenvalues and three eigenvectors of the inertial tensor. The solutions to this reduced nonlinear system are characterized in terms of the intersection curves between variable ellipsoids and the unit sphere: each eigenvector lies on the intersection curve associated with a given eigenvalue.

Following the establishment of the intersection conditions, we derive sharp bounds on a six-dimensional subspace containing all possible inertial tensors of interest, and in particular the inertial tensor of the global minimum. These bounds, together with an explicit parametrization of the intersection curves and the symmetries in the equations, are used to construct a specialized search procedure that systematically samples this subspace in order to extract a set of approximate solutions to the equations. We then refine this set of approximate solutions by means of iterative methods specifically designed for this task. The outcome is, in general, a rather small set of distinct solutions from which the global minimum is found by simply comparing their function values. The entire algorithm is evaluated on a realistic set of simulated and chemical test problems, which demonstrate its effectiveness in solving the above-mentioned conformational problems.

2. Background and problem formulation

We begin with some necessary background on the *EMBED* algorithm. Suppose we have estimates of all the interatomic distances $[d_{ij}]$ in a structure, and we wish to find a matrix of three-dimensional coordinates $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_M]^\top \in \mathbb{R}^{M\times 3}$ for the M atoms such that the distances calculated from the coordinates $\|\mathbf{x}_i - \mathbf{x}_j\|$ in some sense closely match their estimated values. The obvious way to do this is to minimize the weighted *STRESS* function

$$\sum_{1=i<j}^{M,M} \big(w_{ij}\big[\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij}\big]\big)^2, \tag{1}$$

or *SSTRESS* function

$$\sum_{1=i<j}^{M,M} \left(w_{ij}\left[\|\mathbf{x}_i - \mathbf{x}_j\|^2 - D_{ij}\right]\right)^2, \tag{2}$$

with respect to the coordinates, where $D_{ij} \equiv d_{ij}^2$ and $w_{ij} \geqslant 0$ are weights.

Although this can certainly be done [21], it turns out that another function, called the *STRAIN*, has nicer mathematical properties. To explain this, we expand equation (2) as

$$4 \sum_{1=i<j}^{M,M} w_{ij}^2 \left[(\mathbf{x}_i \cdot \mathbf{x}_j) - \left(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - D_{ij}\right)/2\right]^2. \tag{3}$$

By the law of cosines, this can be regarded as four times the weighted sum of the squares of the differences between the dot product $\mathbf{x}_i \cdot \mathbf{x}_j$ and an estimate thereof, namely, $(\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - D_{ij})/2$. This estimate is inconvenient since it depends on the coordinates we are trying to calculate. An estimate that is independent of the coordinates may be obtained from the following formula, which gives the squared distances to the center of mass of the configuration, $M^{-1}\sum_j m_j \mathbf{x}_j$, where $\overline{M} \equiv \sum_j m_j$, in terms of the squared distances among the atoms:

$$D_{0i} = \overline{M}^{-1} \sum_{j=1}^{M} m_j D_{ij} - \overline{M}^{-2} \sum_{1=j<k}^{M,M} m_j m_k D_{jk}. \tag{4}$$

It is straightforward to show that when $D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$ for some set of center of mass coordinates, the estimate D_{0i} is exact, i.e., $D_{0i} = \|\mathbf{x}_i\|^2$ [8]. Moreover, because equation (4) is an averaging procedure, the D_{0i} obtained from it are quite insensitive to errors in the D_{ij}.

The complete matrix of dot products between all pairs of coordinate vectors, or *Gram matrix*, is readily calculated from the coordinates as

$$\mathbf{X}\mathbf{X}^\top = [\mathbf{x}_i \cdot \mathbf{x}_j]_{i,j=1}^{M,M}.$$

If the coordinates $\mathbf{X}$ are center of mass coordinates and

$$\mathbf{D} = [D_{ij}]_{i,j=1}^{M,M} = \left[\|\mathbf{x}_i - \mathbf{x}_j\|^2\right]_{i,j=1}^{M,M},$$

then it can also be written as

$$\mathbf{X}\mathbf{X}^\top = \left[(D_{0i} + D_{0j} - D_{ij})/2\right]_{i,j=1}^{M,M}.$$

Let $\mathbf{I}$ be an identity matrix, $\mathbf{1}$ a column vector of ones, and $\mathbf{m} = [m_1, \ldots, m_M]^\top$. Then this transformation from matrices of squared distances to Gram matrices can be written succinctly (see [13]) as

$$\mathbf{X}\mathbf{X}^\top = -\frac{1}{2}\left[\mathbf{I} - \left(\mathbf{1}\mathbf{m}^\top\right)/\overline{M}\right]\mathbf{D}\left[\mathbf{I} - \left(\mathbf{m}\mathbf{1}^\top\right)/\overline{M}\right]. \tag{5}$$

The mathematical equivalent of the Gram matrix in statistics is the well known covariance matrix.

If the D_{ij} are only inexact estimates of the actual squared distances $\|\mathbf{x}_i - \mathbf{x}_j\|^2$, we can calculate an estimate of the corresponding Gram matrix as

$$\mathbf{A} = [a_{ij}]_{i,j=1}^{M,M} = \big[(D_{0i} + D_{0j} - D_{ij})/2\big]_{i,j=1}^{M,M}.$$

The *STRAIN* may then be defined as

$$\sum_{i,j=1}^{M,M} \big(w_{ij}\big[(\mathbf{x}_i \cdot \mathbf{x}_j) - a_{ij}\big]\big)^2 = \big\|\mathbf{W} \bullet \big[\mathbf{X}\mathbf{X}^\top - \mathbf{A}\big]\big\|_{\mathrm{F}}^2, \tag{6}$$

where $\|\mathbf{C}\|_{\mathrm{F}}^2 = \sum_{ij} c_{ij}^2$ denotes the squared *Frobenius norm* of a matrix **C**, while $\mathbf{W} = [w_{ij}]_{i,j=1}^{M,M}$ is a matrix of weights and the "•" denotes the *Hadamard* (or entry-by-entry) matrix product.

This problem of minimizing the *STRAIN* does not have a simple solution for arbitrary weights, but if $w_{ij} = m_i m_j$ for some assignment of masses to the atoms, equation (6) can be written as

$$\big\|\mathbf{M}\big[\mathbf{X}\mathbf{X}^\top - \mathbf{A}\big]\mathbf{M}\big\|_{\mathrm{F}}^2,$$

where $\mathbf{M} = \mathrm{diag}(\mathbf{m})$ is a nonsingular diagonal matrix containing the masses. The mass-weighted Gram matrix $\mathbf{M}[\mathbf{X}\mathbf{X}^\top]\mathbf{M}$ is also a positive semi-definite rank three matrix. Assuming that the matrix **MAM** also has at least three nonnegative eigenvalues (as is almost always the case in practice), we shall denote its three largest eigenvalues by $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant 0$, and their corresponding eigenvectors by $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$. Then a classical theorem due to Sylvester, Eckart and Young on fixed-rank matrix approximations shows that the *STRAIN* is minimized by the coordinates

$$\mathbf{X}^* = \mathbf{M}^{-1}\big[\lambda_1^{1/2}\mathbf{u}_1, \lambda_2^{1/2}\mathbf{u}_2, \lambda_3^{1/2}\mathbf{u}_3\big]$$

(see [11]). In chemistry, this calculation is known as the *EMBED* algorithm [9]. It should be noted that a variety of reliable iterative methods exist for finding the three largest eigenvalues and associated eigenvectors of the symmetric positive semidefinite matrix **MAM** in far less time than would be needed to fully diagonalize it.

We shall now formulate the main problem of this paper, namely embedding with a rigid substructure. Suppose that we know the $3D$ coordinates of a subset of M atoms **X**, and that we wish to find $3D$ coordinates **Y** for the remaining N atoms so as to minimize the overall *STRAIN*, regarded as a function of **Y** alone. If we order the atoms appropriately, we may partition the Gram matrix into four submatrices and write this function as

$$f(\mathbf{Y}) = \frac{1}{2}\left\|\begin{bmatrix}\mathbf{W}_A & \mathbf{W}_B \\ \mathbf{W}_B^\top & \mathbf{W}_c\end{bmatrix} \bullet \left(\begin{bmatrix}\mathbf{X}\mathbf{X}^\top & \mathbf{X}\mathbf{Y}^\top \\ \mathbf{Y}\mathbf{X}^\top & \mathbf{Y}\mathbf{Y}^\top\end{bmatrix} - \begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{C}\end{bmatrix}\right)\right\|_{\mathrm{F}}^2, \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{M\times M}$, $\mathbf{B} \in \mathbb{R}^{M\times N}$, $\mathbf{C} \in \mathbb{R}^{N\times N}$, while the matrices of weights satisfy $\mathbf{W}_A, \mathbf{W}_B, \mathbf{W}_C \geqslant \mathbf{0}$. Since $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$ by definition, equation (7) may be expanded and simplified as

$$f(\mathbf{Y}) = \big\|\mathbf{W}_B \bullet \big[\mathbf{X}\mathbf{Y}^\top - \mathbf{B}\big]\big\|_{\mathrm{F}}^2 + \frac{1}{2}\big\|\mathbf{W}_C \bullet \big[\mathbf{Y}\mathbf{Y}^\top - \mathbf{C}\big]\big\|_{\mathrm{F}}^2, \tag{8}$$

which shows that $\mathbf{W}_A$ is irrelevant.

Before turning to the development of methods for finding the global minimum of the *STRAIN* function in equation (8), we list two assumptions. First, the entries in $\mathbf{Y}$ are independent. Second, we shall consider only the unweighted problem, i.e., $w_{ij} = 1$ in both $\mathbf{W}_B$ and $\mathbf{W}_C$. It should be straightforward to extend our methods to the special case of multiplicative weights satisfying $w_{ij} = m_i m_j$, which may prove useful in some applications. The problem involving general weights will not be addressed in this paper.

3. First-order optimality conditions

3.1. The nonlinear matrix equation

We begin by deriving the first-order optimality conditions for the minimization of the unweighted *STRAIN* function

$$\min_{\mathbf{Y}} \big(f(\mathbf{Y})\big) = \min_{\mathbf{Y}} \Big(\big\|\mathbf{X}\mathbf{Y}^\top - \mathbf{B}\big\|_{\mathrm{F}}^2 + \frac{1}{2}\big\|\mathbf{Y}\mathbf{Y}^\top - \mathbf{C}\big\|_{\mathrm{F}}^2\Big). \tag{9}$$

In the following, we let $\mathbf{G}_\mathbf{Y} = \mathbf{X}^\top\mathbf{X} + \mathbf{Y}^\top\mathbf{Y} \in \mathbb{R}^{3\times 3}$, so that $\mathbf{G}_\mathbf{0} = \mathbf{X}^\top\mathbf{X}$, and $\mathbf{V} = \mathbf{B}^\top\mathbf{X} \in \mathbb{R}^{N\times 3}$. Because $\mathbf{G}_\mathbf{Y}$ is simply related to the usual inertial tensor, i.e., $\mathrm{tr}(\mathbf{G}_\mathbf{Y})\mathbf{I}_3 - \mathbf{G}_\mathbf{Y}$, we shall often refer to it as the inertial tensor. A matrix $\mathbf{Y}_c$ at which the gradient $\nabla f(\mathbf{Y}_c) = \mathbf{0}$ shall be called a *critical matrix*.

Theorem 1. A critical matrix satisfies the nonlinear matrix equation

$$\mathbf{Y}\mathbf{G}_\mathbf{Y} - \mathbf{C}\mathbf{Y} - \mathbf{V} = \mathbf{0}. \tag{10}$$

Among all critical matrices $\mathbf{Y}_c$, the one which maximizes the quantity

$$\mathrm{tr}\big(\mathbf{Y}_c^\top\mathbf{V}\big) + \frac{1}{2}\,\mathrm{tr}\big(\big[\mathbf{Y}_c^\top\mathbf{Y}_c\big]^2\big) \tag{11}$$

is the global minimum of f.

Proof. Expanding equation (9) yields

$$f(\mathbf{Y}) = \mathrm{tr}\Big(\big[\mathbf{X}\mathbf{Y}^\top - \mathbf{B}\big]^\top\big[\mathbf{X}\mathbf{Y}^\top - \mathbf{B}\big]\Big) + \frac{1}{2}\,\mathrm{tr}\Big(\big[\mathbf{Y}\mathbf{Y}^\top - \mathbf{C}\big]^\top\big[\mathbf{Y}\mathbf{Y}^\top - \mathbf{C}\big]\Big)$$

$$= \mathrm{tr}\Big(\mathbf{B}^\top\mathbf{B} + \frac{1}{2}\mathbf{C}^2\Big) - \mathrm{tr}\big([2\mathbf{V} + \mathbf{CY}]\mathbf{Y}^\top\big) + \mathrm{tr}\Big(\mathbf{Y}\mathbf{G}_\mathbf{Y}\mathbf{Y}^\top - \frac{1}{2}\mathbf{Y}\mathbf{Y}^\top\mathbf{Y}\mathbf{Y}^\top\Big). \quad (12)$$

The gradient $\nabla f(\mathbf{Y})$ can be arranged in the same matrix form as $\mathbf{Y}$, namely,

$$\nabla f(\mathbf{Y}) \equiv \frac{\partial f}{\partial \mathbf{Y}} \equiv \begin{bmatrix} \partial f/\partial y_{11} & \partial f/\partial y_{12} & \partial f/\partial y_{13} \\ \cdots & \cdots & \cdots \\ \partial f/\partial y_{N1} & \partial f/\partial y_{N2} & \partial f/\partial y_{N3} \end{bmatrix}, \quad (13)$$

which facilitates the calculation of the gradients of matrix functions [14]. In particular, the matrix of $\nabla\,\mathrm{tr}(\mathbf{M}^\top\mathbf{N})$ with respect to $\mathbf{M}$ is

$$\frac{\partial\,\mathrm{tr}(\mathbf{M}^\top\mathbf{N})}{\partial\mathbf{M}} = \frac{\partial\,\mathrm{tr}(\mathbf{N}^\top\mathbf{M})}{\partial\mathbf{M}} = \mathbf{N} \quad (14)$$

for any two conformable matrices $\mathbf{M}$ and $\mathbf{N}$. If we differentiate each term in equation (12), using both equation (14) and the rule for differentiating a matrix product [14], and then simplify using the invariance of the trace of a matrix product under cyclic permutations, we obtain equation (10).

Given a solution $\mathbf{Y}_c$ of equation (10), we may replace $\mathbf{Y}\mathbf{G}_\mathbf{Y}$ in the last term of equation (12) with $\mathbf{CY}_c + \mathbf{V}$ and simplify to get

$$f(\mathbf{Y}_c) = \mathrm{tr}\Big(\mathbf{B}^\top\mathbf{B} + \frac{1}{2}\mathbf{C}^2\Big) - \mathrm{tr}\big(\mathbf{Y}_c^\top\mathbf{V}\big) - \frac{1}{2}\,\mathrm{tr}\big(\big[\mathbf{Y}_c^\top\mathbf{Y}_c\big]^2\big), \quad (15)$$

thus showing why the quantity in equation (11) must be maximized. □

Two comments are in order. First, because $\mathbf{G}_\mathbf{Y} = \mathbf{X}^\top\mathbf{X} + \mathbf{Y}^\top\mathbf{Y}$, equation (10) constitutes a system of $3N$ cubic equations in the entries of $\mathbf{Y}$. The direct solution of equation (10) does not appear theoretically or computationally feasible, and alternative approaches are essential if the global minimum is to be found. Second, if there are no fixed coordinates, then equation (10) becomes

$$\mathbf{CY} = \mathbf{Y}\mathbf{G}_\mathbf{Y}. \quad (16)$$

The matrix $\mathbf{G}_\mathbf{Y} = \mathbf{Y}^\top\mathbf{Y}$ is a 3×3 symmetric positive-definite matrix whose spectral decomposition is denoted by $\mathbf{G}_\mathbf{Y} = \mathbf{R}\Gamma\mathbf{R}^\top$, γ_k being the kth diagonal entry in Γ and $k = 1, 2, 3$. If one defines $\widehat{\mathbf{Y}} = \mathbf{YR}$, so that $\widehat{\mathbf{y}}_k$ is its kth column, then equation (16) becomes $\mathbf{C}\widehat{\mathbf{Y}} = \widehat{\mathbf{Y}}\Gamma$ or $\mathbf{C}\widehat{\mathbf{y}}_k = \widehat{\mathbf{y}}_k\gamma_k$. Thus each $\widehat{\mathbf{y}}_k$ is proportional to an eigenvector of $\mathbf{C}$, and since $\Gamma = \widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}$ the constants of proportionality must be $\sqrt{\gamma_k}$. Finally, $f(\mathbf{Y}) = f(\widehat{\mathbf{Y}})$ is minimized when

$$\mathrm{tr}\big(\big[\widehat{\mathbf{Y}}^\top\widehat{\mathbf{Y}}\big]^2\big) = \sum_k \gamma_k^2$$

is maximized. This is the theorem of Sylvester, Eckart and Young mentioned in section 2.

3.2. Equations for the inertial tensor

In this section we derive a new system of equations and show its relation to the matrix equation (10). The result is that the number of variables needed to solve the minimization problem is dramatically reduced from $3N$ to 6. We start from the spectral factorization of the 3×3 inertial tensor $\mathbf{G_Y}$,

$$\mathbf{G_Y} = \mathbf{G_0} + \mathbf{Y}^\top\mathbf{Y} = \mathbf{R}\Gamma\mathbf{R}^\top, \tag{17}$$

order the diagonal entries of Γ as $\gamma_1 \geqslant \gamma_2 \geqslant \gamma_3 > 0$ and, if necessary, change the signs of the columns of $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$ so that it is a proper orthogonal matrix. The twelve unknown entries in Γ and $\mathbf{R}$ will be the main variables throughout the rest of this paper. It turns out that a few largest eigenvalues of the symmetric matrix $\mathbf{C}$ play a significant role, so we will write its spectral factorization as $\mathbf{C} = \mathbf{Q}\Sigma\mathbf{Q}^\top$ with the diagonal entries of Σ ordered as $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_N$ while $\mathbf{Q}$ is an orthogonal matrix. We shall use standard notation for the Kronecker delta δ_{kl}, and $\mathbf{I}$ (or $\mathbf{I}_N$) for the $N \times N$ identity matrix. We now define an $N \times 3$ matrix $\mathbf{W} = \mathbf{Q}^\top\mathbf{V}$ whose ith row is denoted by $\mathbf{w}_i^\top$, and two kinds of 3×3 variable symmetric matrices that will frequently occur, namely, a univariate one,

$$\mathbf{S}(\gamma) = \mathbf{G_0} + \mathbf{W}^\top[\gamma\mathbf{I} - \Sigma]^{-2}\mathbf{W} = \mathbf{G_0} + \sum_{i=1}^{N} \frac{\mathbf{w}_i\mathbf{w}_i^\top}{(\gamma - \sigma_i)^2}, \tag{18}$$

and a bivariate one

$$\begin{aligned}\mathbf{S}(\gamma, \delta) &= \mathbf{G_0} + \mathbf{W}^\top\big([\gamma\mathbf{I} - \Sigma]^{-1}[\delta\mathbf{I} - \Sigma]^{-1}\big)\mathbf{W} \\ &= \mathbf{G_0} + \sum_{i=1}^{N} \frac{\mathbf{w}_i\mathbf{w}_i^\top}{(\gamma - \sigma_i)(\delta - \sigma_i)}.\end{aligned} \tag{19}$$

Theorem 2. If $\mathbf{Y}_c$ is a critical matrix of f, then the eigenvalues and eigenvectors $\{\gamma_1, \mathbf{r}_1; \gamma_2, \mathbf{r}_2; \gamma_3, \mathbf{r}_3\}$ of the associated inertial tensor $\mathbf{G}_{\mathbf{Y}_c}$ satisfy the system of six equations

$$\begin{aligned}\gamma_k - \mathbf{r}_k^\top\mathbf{S}(\gamma_k)\mathbf{r}_k = 0, \quad k = 1, 2, 3, \\ \mathbf{r}_k^\top\mathbf{S}(\gamma_k, \gamma_l)\mathbf{r}_l = 0, \quad 1 \leqslant k < l \leqslant 3,\end{aligned} \tag{20}$$

as well as the six orthonormality conditions

$$\mathbf{r}_k^\top\mathbf{r}_l = \delta_{kl}. \tag{21}$$

Proof. The orthonormality relations (21) hold for the eigenvectors of any symmetric matrix. By premultiplying each eigenvector equation $\mathbf{G_Y}\mathbf{r}_k = \gamma_k\mathbf{r}_k$ by $\mathbf{r}_l^\top$, we observe that there are only six distinct such products for $1 \leqslant k \leqslant l \leqslant 3$. These products are equivalent, by equations (17) and (21), to the following system of six equations:

$$\delta_{kl}\gamma_k = \mathbf{r}_k^\top\mathbf{G_Y}\mathbf{r}_l = \mathbf{r}_k^\top\mathbf{G_0}\mathbf{r}_l + (\mathbf{Y}\mathbf{r}_k)^\top(\mathbf{Y}\mathbf{r}_l). \tag{22}$$

To obtain an expression for $\mathbf{Y}_c\mathbf{r}_k$ we multiply equation (10) with the matrix $\mathbf{R}$ and substitute equation (17) for $\mathbf{G_Y}$. The result is $\mathbf{Y}_c\mathbf{R}\Gamma = \mathbf{C}\mathbf{Y}_c\mathbf{R}+\mathbf{V}\mathbf{R}$ which is equivalent to the three vector equations $(\gamma_k\mathbf{I}-\mathbf{C})\mathbf{Y}_c\mathbf{r}_k = \mathbf{V}\mathbf{r}_k$, one for each column of $\mathbf{R}$, whose formal solutions are given by

$$\mathbf{Y}_c\mathbf{r}_k = [\gamma_k\mathbf{I}-\mathbf{C}]^{-1}\mathbf{V}\mathbf{r}_k, \quad k = 1,2,3. \tag{23}$$

Substitution of equation (23) into equation (22) and rearrangement results in

$$\delta_{kl}\gamma_k - \mathbf{r}_k^\top\big[\mathbf{G_0} + \mathbf{V}^\top\big([\gamma_k\mathbf{I}-\mathbf{C}]^{-1}[\gamma_l\mathbf{I}-\mathbf{C}]^{-1}\big)\mathbf{V}\big]\mathbf{r}_l = 0. \tag{24}$$

Separation of these equations into the $k = l$ and $k < l$ cases, and substitution of $\mathbf{C} = \mathbf{Q}\Sigma\mathbf{Q}^\top$ and $\mathbf{V} = \mathbf{Q}\mathbf{W}$, now yields equations (20). □

Equations (20) and (21) can be regarded as a system of twelve equations in the twelve unknowns $\{\gamma_1, \mathbf{r}_1; \gamma_2, \mathbf{r}_2; \gamma_3, \mathbf{r}_3\}$. The six constraints in equation (21) on the nine entries of $\mathbf{R}$ can be eliminated by a suitable parametrization for the 3×3 orthogonal matrices. Two specific parametrizations will be introduced later on, the outcomes being two different nonlinear systems of six equations in six unknowns. We shall often refer to equations (20) as the *inertial equations*, and denote them by g_{kl}, for $1 \leqslant k \leqslant l \leqslant 3$. An eigenvalue–eigenvector pair $\{\gamma_k, \mathbf{r}_k\}$ will be referred to as an *inertial pair*. Each of the first three equations depends on a single inertial pair and will be referred to as *quadratic*, while the last three equations, each depending on two inertial pairs, will be referred to as *bilinear*.

The connection between inertial pairs satisfying equations (20)–(21) and critical matrices $\mathbf{Y}_c$ is that given any positive diagonal matrix Γ, such that each diagonal entry satisfies $0 < \gamma_k \neq \sigma_i$, and an arbitrary 3×3 orthogonal matrix $\mathbf{R}$, one can compute the matrix $\mathbf{R}\Gamma\mathbf{R}^\top$ and the matrix

$$\mathbf{Y} = \big[(\gamma_1\mathbf{I}-\mathbf{C})^{-1}\mathbf{V}\mathbf{r}_1 \mid (\gamma_2\mathbf{I}-\mathbf{C})^{-1}\mathbf{V}\mathbf{r}_2 \mid (\gamma_3\mathbf{I}-\mathbf{C})^{-1}\mathbf{V}\mathbf{r}_3\big]\mathbf{R}^\top \tag{25}$$

satisfying equation (10). In general, however, $\mathbf{R}\Gamma\mathbf{R}^\top \neq \mathbf{G_0} + \mathbf{Y}^\top\mathbf{Y}$ unless the matrix pair $\{\Gamma, \mathbf{R}\}$ is also a solution to equations (20), in which case $\mathbf{Y}$ is a critical matrix of f. The same critical matrix is obtained for each of the eight possible choices of signs $\pm\mathbf{r}_k$ $(k = 1,2,3)$. The choice of a particular orientation will be described in section 5.2. In the following, after computing solutions to equations (20)–(21), the corresponding critical matrix will be constructed according to equation (25).

4. The structure of the solution space

4.1. The intersection curve between a sphere and an ellipsoid

Let $g(\gamma, \mathbf{r}) = 0$ denote any one of the three quadratic inertial equations in (20), with the associated matrix $\mathbf{S}(\gamma)$ defined by equation (18), and let $\{\gamma, \mathbf{r}\}$ denote an inertial pair that satisfies it. A geometric characterization of all such inertial pairs comes from the following result:

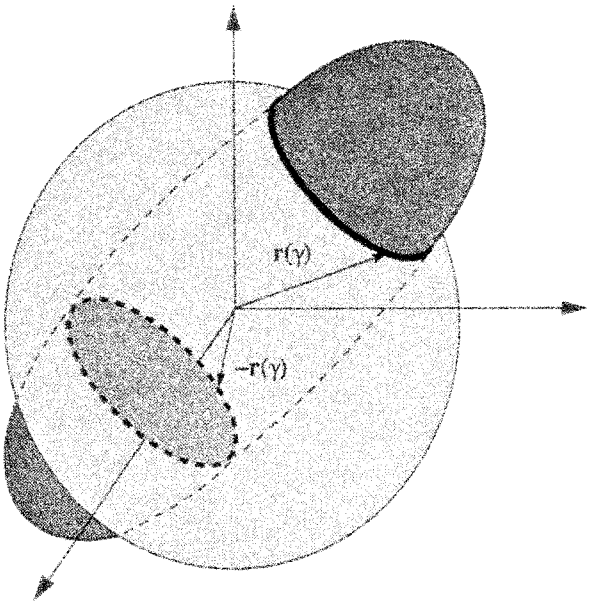

Figure 1. The two symmetric curves in which the ellipsoid defined by the matrix $\mathbf{S}(\gamma)/\gamma$ intersects the unit sphere.

Lemma 3. If an inertial pair $\{\gamma, \mathbf{r}\}$ satisfies the quadratic inertial equation $g(\gamma, \mathbf{r}) = 0$, then $\mathbf{r}$ is located on the intersection curve between the unit sphere and the ellipsoid defined by the matrix $\mathbf{S}(\gamma)/\gamma$.

Proof. Since $\gamma > 0$ the equation $g(\gamma, \mathbf{r}) = 0$ can be rewritten as

$$\mathbf{r}^\top \left[\frac{\mathbf{S}(\gamma)}{\gamma} \right] \mathbf{r} = 1. \tag{26}$$

For any fixed $0 < \gamma \neq \sigma_i$, the matrix $\mathbf{S}(\gamma)/\gamma$ is positive-definite by (18), and hence any vector $\mathbf{r}$ that satisfies equation (26) lies on the surface of an ellipsoid centered at the origin. As a column of an orthogonal matrix, however, the vector $\mathbf{r}$ also lies on the surface of the unit sphere centered at origin. Therefore, $\mathbf{r}$ lies on the intersection curve between the unit sphere and the ellipsoid defined by γ via equation (26) (see figure 1). □

It is well known that the spectral factorization of $\mathbf{S}(\gamma)/\gamma = \mathbf{U}\Lambda\mathbf{U}^\top$ determines the principal axes of this ellipsoid as the three eigenvectors in $\mathbf{U}$, while the three eigenvalues $0 < \lambda_3 \leqslant \lambda_2 \leqslant \lambda_1$ are the reciprocal squares of the three semiaxes. It is also evident from equations (18) and (26) that both $\Lambda = \Lambda(\gamma)$ and $\mathbf{U} = \mathbf{U}(\gamma)$ are continuous functions, except at the eigenvalues of $\mathbf{C}$ (see also [19]). Therefore, the problem of existence of an intersection curve, henceforth denoted by $\mathbf{r}(\gamma)$, consists of finding intervals of γ over which the quadratic vector equation (26) has real solutions.

Since the ellipsoid and the unit sphere are centrally symmetric convex surfaces, the necessary and sufficient condition for their intersection is

$$\lambda_3(\gamma) \leqslant 1 \leqslant \lambda_1(\gamma). \tag{27}$$

In other words, there is no intersection if all semiaxes are strictly greater, or strictly smaller, than the radius of the sphere.

By the central symmetry of both surfaces, the intersection consists of a symmetrical pair of (nonintersecting) closed spherical curves, related to each other by reflection in the origin. The curves are symmetric with respect to all three axes, but we shall be mostly interested in the axis of symmetry along which piercing of surfaces occurs: $\mathbf{u}_1$ if $\lambda_2(\gamma) < 1$, or $\mathbf{u}_3$ if $1 < \lambda_2(\gamma)$. We shall label this axis vector by $\mathbf{u}_k$, and the other two axes by $\mathbf{u}_i$ and $\mathbf{u}_j$, so that $k = 1$ implies $(i, j) = (2, 3)$ while $k = 3$ implies $(i, j) = (1, 2)$. At a *bifurcation point* γ_b, which is defined by the nonlinear equation

$$\lambda_2(\gamma_b) = 1, \tag{28}$$

the two intersection curves merge at the points $\pm\mathbf{u}_2$ while the axis of symmetry for both curves undergoes a discontinuous change to $\mathbf{u}_2$.

The following lemma shows how the explicit analytic parametrization of the intersection curve depends on the axis of symmetry.

Lemma 4. The projection of both intersection curves onto the central plane orthogonal to the axis of symmetry $\mathbf{u}_k$ is an ellipse, whose semiaxes are given by

$$\begin{aligned} \beta_i(\gamma) &= \sqrt{(\lambda_k - 1)/(\lambda_k - \lambda_i)}, \\ \beta_j(\gamma) &= \sqrt{(\lambda_k - 1)/(\lambda_k - \lambda_j)}. \end{aligned} \tag{29}$$

Any regular parametrization of an ellipse can be chosen to parametrize the intersection curve. We shall use the trigonometric one given by

$$\begin{aligned} \xi_i &= \beta_i(\gamma)\cos(\psi), \\ \xi_j &= \beta_j(\gamma)\sin(\psi), \\ \xi_k &= \pm\sqrt{1 - \xi_i^2 - \xi_j^2}, \end{aligned} \tag{30}$$

for $\psi \in [0, 2\pi)$. The vector $\xi(\gamma, \psi) = [\xi_1(\gamma, \psi), \xi_2(\gamma, \psi), \xi_3(\gamma, \psi)]^\top$ is the representation of the intersection curve with respect to the principal axes, while its representation with respect to the standard axes is given by

$$\mathbf{r}(\gamma, \psi) = \mathbf{U}(\gamma)\xi(\gamma, \psi). \tag{31}$$

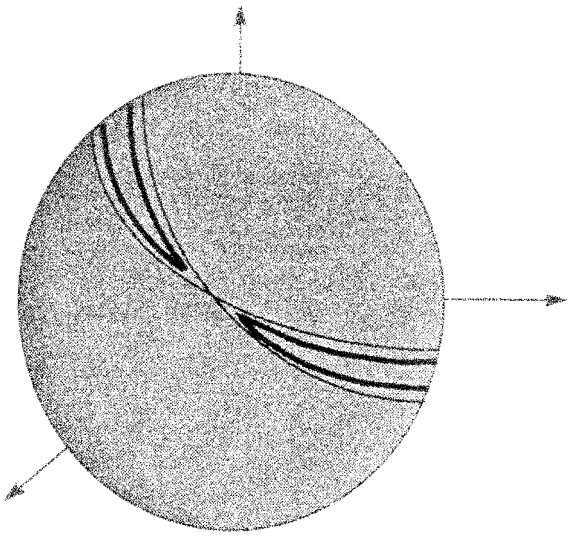

Figure 2. The shape of the intersection curves due to the narrow ellipsoids near a bifurcation point.

Proof. If we write the equations for the unit sphere and the ellipsoid in the principal axis coordinates ξ of the ellipsoid, and then eliminate the coordinate ξ_k from these equations, we get

$$\left(\frac{\lambda_k - \lambda_i}{\lambda_k - 1}\right)\xi_i^2 + \left(\frac{\lambda_k - \lambda_j}{\lambda_k - 1}\right)\xi_j^2 = 1. \tag{32}$$

Equation (32) describes an ellipse in the plane spanned by the ith and jth eigenvector, whose associated semiaxes are given by equation (29). If we choose an angle ψ for the trigonometric parametrization of the ellipse, and back substitute for the eliminated coordinate ξ_k, we get equation (30). The vector $\xi(\gamma, \psi)$ refers to the principal axes; transformation to the original coordinates comes from the identification $\mathbf{U}^\top \mathbf{r} = \xi(\gamma, \psi)$, which proves equation (31). □

Lemma 4 provides a simple and efficient way to compute any point on any intersection curve where, according to lemma 3, all solutions to the inertial equations are to be found. Since we shall also require the derivatives of $\mathbf{r}(\gamma, \psi)$ with respect to both variables during the search for the global minimum, we should analyze the regions where these derivatives are ill-behaved in order to prevent difficulties or even failure of the search. Geometrically this happens if a solution is located on a set of very small, or very narrow, intersection curves. Analytically, the small intersection curves occur when both projected semiaxes $\beta_i(\gamma), \beta_j(\gamma) \to 0$ simultaneously, which happens when either $\lambda_1(\gamma) \to 1$ or $\lambda_3(\gamma) \to 1$. Narrow intersection curves occur when the projected eccentricity $\min((\beta_i/\beta_j), (\beta_j/\beta_i)) \ll 1$. Almost all such cases occur in the vicinity of those bifurcation points that have a very small angle between the intersection curves (see figure 2).

To compute this angle α, we eliminate ξ_2 from the equations for an ellipsoid and the unit sphere, set $\lambda_2(\gamma_b) = 1$, and obtain

$$\tan\Big(\frac{\alpha}{2}\Big) = \frac{\xi_1}{\xi_3} = \pm\sqrt{\frac{1-\lambda_3(\gamma_b)}{\lambda_1(\gamma_b)-1}}. \tag{33}$$

This shows that the very small angles are caused by either $\lambda_3(\gamma_b)$ being very close to 1 (which is also a near-tangency case), or by $\lambda_1(\gamma_b) \gg 1$ (which happens when a bifurcation point is very close to an eigenvalue σ_k), or both.

4.2. Spherical bounds on the moments of inertia

Let $\{\gamma, \mathbf{r}\}$ be an unknown inertial pair satisfying the quadratic equation $g(\gamma, \mathbf{r}) = 0$. Since $\mathbf{r}$ is confined to the unit sphere, it is of great interest to find an upper bound on $\gamma > 0$ as well as a better lower bound, because such bounds would reduce the size of the subspace over which the search for the global minimum takes place. The notation $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ stands for the maximal and minimal eigenvalue of the argument matrix (e.g., from equation (26), $\lambda_{\max}(\mathbf{S}(\gamma)) = \gamma\lambda_1(\gamma)$).

Theorem 5. Let γ^+ be the maximal solution of the nonlinear equation

$$\gamma = \lambda_{\max}\big(\mathbf{S}(\gamma)\big). \tag{34}$$

Then the largest eigenvalue γ_1 of the optimal inertial tensor is bounded by

$$\max\big(\sigma_1, \lambda_{\max}(\mathbf{G_0})\big) \leqslant \gamma_1 \leqslant \gamma^+. \tag{35}$$

Proof. We start by writing the quadratic inertial equation (26) as $\gamma = \mathbf{r}^\top \mathbf{S}(\gamma)\mathbf{r}$. The maximum of its right-hand side over all unit vectors $\mathbf{r}$, (also known as the Rayleigh quotient of $\mathbf{S}(\gamma)$) is $\lambda_{\max}(\mathbf{S}(\gamma))$ (see, e.g., [3]), and the resulting inequality

$$\gamma \leqslant \max_{\|\mathbf{r}\|=1} \big(\mathbf{r}^\top \mathbf{S}(\gamma)\mathbf{r}\big) = \lambda_{\max}\big(\mathbf{S}(\gamma)\big), \tag{36}$$

yields the nonlinear equation (34). If the maximal positive solution of equation (34) exists, then it provides an upper bound on any eigenvalue γ, and in particular on the largest one γ_1.

To prove the existence of a solution to equation (34), we recall that $\mathbf{S}(\gamma)$ is a positive-definite matrix, so that the spectral radius function

$$\rho(\gamma) = \lambda_{\max}\big(\mathbf{S}(\gamma)\big) = \lambda_{\max}\left(\mathbf{G_0} + \sum_{i=1}^{N} \frac{\mathbf{w}_i \mathbf{w}_i^\top}{(\gamma - \sigma_i)^2}\right) \tag{37}$$

is strictly positive. It is also a rational function with poles of order two at the eigenvalues of $\mathbf{C}$. In the interval (σ_1, ∞), it is convex and decreases monotonically from infinity at the pole σ_1 to the asymptotic value of $\lambda_{\max}(\mathbf{G_0})$. We omit the proofs of these statements, which are based on the negative and positive semi-definiteness of the

first and second derivatives of $\mathbf{S}(\gamma)$, respectively. It follows that in the interval (σ_1, ∞) there is only one solution, denoted by γ^+, to the equation $\gamma = \rho(\gamma)$, which is also the maximal solution to the inequality $\gamma \leqslant \rho(\gamma)$.

To prove the lower bound on γ_1 in equation (35), consider first the case when $\sigma_1 \leqslant \lambda_{\max}(\mathbf{G_0})$. Then, rewriting equation (17) as $\Gamma = \mathbf{R}^\top(\mathbf{G_0} + \mathbf{Y}^\top\mathbf{Y})\mathbf{R}$, we see that $\gamma_1 = \lambda_{\max}((\mathbf{R}^\top(\mathbf{G_0} + \mathbf{Y}^\top\mathbf{Y})\mathbf{R})) = \lambda_{\max}(\mathbf{G_0} + \mathbf{Y}^\top\mathbf{Y}) \geqslant \lambda_{\max}(\mathbf{G_0})$, by the positive semi-definiteness of $\mathbf{Y}^\top\mathbf{Y}$ and the positive definiteness of $\mathbf{G_0}$. In the case when $\lambda_{\max}(\mathbf{G_0}) < \sigma_1$, let $\{\gamma_1, \mathbf{r}_1\}$ be the first inertial pair satisfying the quadratic equation (26) and proceed as follows:

$$\gamma_1 = \mathbf{r}_1^\top \left(\mathbf{G_0} + \sum_{i=1}^{N} \frac{\mathbf{w}_i \mathbf{w}_i^\top}{(\gamma_1 - \sigma_i)^2} \right) \mathbf{r}_1 \geqslant \mathbf{r}_1^\top \mathbf{G_0} \mathbf{r}_1 + \frac{(\mathbf{r}_1^\top \mathbf{w}_1)^2}{(\gamma_1 - \sigma_1)^2}$$
$$\geqslant \lambda_{\min}(\mathbf{G_0}) + \frac{(\mathbf{r}_1^\top \mathbf{w}_1)^2}{(\gamma_1 - \sigma_1)^2}. \tag{38}$$

The first inequality comes from keeping only the first term under the sum (all terms are nonnegative), while the second one comes from minimizing the Rayleigh quotient of $\mathbf{G_0}$. By rewriting the final inequality in (38) as

$$(\gamma_1 - \sigma_1)^3 + \big(\sigma_1 - \lambda_{\min}(\mathbf{G_0})\big)(\gamma - \sigma_1)^2 \geqslant \big(\mathbf{r}_1^\top \mathbf{w}_1\big)^2$$

and defining $\delta = \gamma_1 - \sigma_1$, we obtain the cubic inequality

$$\delta^3 + \big(\sigma_1 - \lambda_{\min}(\mathbf{G_0})\big)\delta^2 - \big(\mathbf{r}_1^\top \mathbf{w}_1\big)^2 \geqslant 0. \tag{39}$$

There is always one nonnegative solution to the equality in (39) (see discussion following equation (42)), which means that $\gamma_1 \geqslant \sigma_1$. This would complete the proof for the lower bound in equation (35), were it not for occasional occurrence of two negative real solutions to the inequality (39). We claim that they can be ignored since any value between them does not lead to a *global* minimum. □

The inequality (35) says that there is no intersection between the variable ellipsoid and the unit sphere as long as $\gamma > \lambda_{\max}(\mathbf{S}(\gamma))$. The bound γ^+ is the tangency point at which intersection commences for the first time, and so we have named it the "spherical bound".

We shall solve equation (34) by a rational interpolation method which consists of two iterations. An outer iteration calculates the coefficients of a rational interpolant to the function $\rho(\gamma)$ defined by equation (37), and monitors convergence to the solution of equation (34), while an inner one computes the maximal root of a special cubic equation similar to (39). The outer iteration interpolates $\rho(\gamma)$ by the simplest rational function having a pole of order two at σ_1, namely,

$$\widehat{\rho}(\gamma) = c + \frac{d}{(\gamma - \sigma_1)^2}. \tag{40}$$

The coefficients c, d are computed from the interpolatory data $\rho^{i-1} = \rho(\gamma^{i-1})$ and $\rho^i = \rho(\gamma^i)$ at the last two iterates γ^{i-1} and γ^i, as

$$d = \frac{\rho^i - \rho^{i-1}}{(\gamma^i - \sigma_1)^{-2} - (\gamma^{i-1} - \sigma_1)^{-2}}, \qquad c = \rho^i - \frac{d}{(\gamma^i - \sigma_1)^2}. \tag{41}$$

The next iterate γ^{i+1} is obtained by solving the equation $\gamma = \widehat{\rho}(\gamma)$ to working precision, which is the task of the inner iteration. Given the current values of the coefficients c, d, and with the change of variable $\delta = \gamma - \sigma_1$, this equation takes on the following form:

$$\delta^3 + (\sigma_1 - c)\delta^2 - d = 0. \tag{42}$$

Since $\rho(\gamma)$ is a strictly decreasing function for $\gamma > \sigma_1$, it follows from equation (41) that $d > 0$ for all iterates $\gamma^{i-1}, \gamma^i > \sigma_1$. Therefore, the cubic polynomial in equation (42) is strictly negative at $\delta = 0$, which guarantees that it has a single positive root. A simple transformation of equation (42) yields the following bounds on this root:

$$\begin{aligned} &\sqrt{\frac{d}{\sigma_1 - c + \sqrt{d/(\sigma_1 - c)}}} < \delta < \sqrt{\frac{d}{\sigma_1 - c}} && \text{if } \sigma_1 - c > 0, \\ &|\sigma_1 - c| < \delta < |\sigma_1 - c| + \frac{d}{(\sigma_1 - c)^2} && \text{if } \sigma_1 - c < 0. \end{aligned} \tag{43}$$

For any point in the interval (43), both the first and the second derivative of the cubic polynomial in equation (42) are positive. This means that Newton's method converges unconditionally and monotonically, except possibly for the first step, to the positive root δ^i of equation (42). This completes the inner iteration, after which we set $\gamma^{i+1} = \sigma_1 + \delta^i$ in the outer one.

By using the convexity of $\rho(\gamma)$ in the interval (σ_1, ∞), it can be shown that this rational interpolation procedure generates a sequence γ^i which converges superlinearly to a solution of $\rho(\gamma) = \gamma$ starting from any pair of estimates $\gamma^1, \gamma^2 > \sigma_1$ [20]. In practice, the convergence rate tends to be nearly quadratic since $\widehat{\rho}(\gamma)$ inherits the convexity of $\rho(\gamma)$.

A good starting point for the outer iteration is obtained by computing an upper bound on the right-hand side of equation (37) as

$$\begin{aligned} \rho(\gamma) &\leqslant \lambda_{\max}(\mathbf{G_0}) + \lambda_{\max}\left(\sum_{i=1}^{N} \frac{\mathbf{w}_i \mathbf{w}_i^\top}{(\gamma - \sigma_i)^2}\right) \\ &\leqslant \lambda_{\max}(\mathbf{G_0}) + \sum_{i=1}^{N} \frac{\mathbf{w}_i^\top \mathbf{w}_i}{(\gamma - \sigma_i)^2} \leqslant \lambda_{\max}(\mathbf{G_0}) + \frac{\|\mathbf{W}\|_{\mathrm{F}}^2}{(\gamma - \sigma_1)^2}. \end{aligned} \tag{44}$$

The last inequality is valid for all $\gamma \in (\sigma_1, \infty)$. On cross-multiplying and setting $\delta = \gamma - \sigma_1$, we obtain a cubic inequality

$$\delta^3 + \big(\sigma_1 - \lambda_{\max}(\mathbf{G_0})\big)\delta^2 - \|\mathbf{W}\|_{\mathrm{F}}^2 \leqslant 0, \tag{45}$$

quite similar to equations (42) and (39). The maximum real solution δ_0 to equation (45) is strictly positive, so that $\gamma_1 \leqslant \gamma^+ \leqslant \sigma_1 + \delta_0$. The upper bound $\sigma_1 + \delta_0$ is generally already quite good and, together with any other point in the interval $(\sigma_1, \sigma_1 + \delta_0)$, provides the desired starting pair for the iteration in equations (40)–(41).

A lower bound γ^- on all the eigenvalues γ is obtained by minimizing the Rayleigh quotient in $\gamma = \mathbf{r}^\top \mathbf{S}(\gamma)\mathbf{r}$. The result is

$$\gamma \geqslant \lambda_{\min}\big(\mathbf{S}(\gamma)\big) = \mu(\gamma). \tag{46}$$

We use a rational interpolation procedure, similar to that in equations (40)–(41) to solve for the minimum value γ^- such that $\mu(\gamma^-) = \gamma^-$. Knowing that $\gamma^- \geqslant \lambda_{\min}(\mathbf{G}_0)$, we start the search for eigenvalues other than γ_1 from $\lambda_{\min}(\mathbf{G}_0)$. The meaning of equation (46) is that the ellipsoid $\mathbf{S}(\gamma)/\gamma$ is guaranteed to have all its semiaxes less than 1 for any $\gamma < \gamma^-$, so there are no further intersections with the unit sphere.

4.3. Conical bounds on the moments of inertia

By definition, there are two antipodal tangency points at $\gamma = \gamma^+$, where the ellipsoid touches the unit sphere contained within it. As γ starts decreasing from γ^+ towards σ_1, so does the size of the ellipsoid, because all three semiaxes decrease. The two intersection curves start growing until either the pole σ_1 or a bifurcation point γ_b is reached, if there is one in this interval. At the bifurcation point the intersection curves meet at a single pair of antipodal points. On further decreasing the value of γ, the curves bifurcate (switch to complementary regions defined by equation (33), see figure 2), but now begin to shrink in size. An important property of the continuous variations of the intersection curves $\mathbf{r}(\gamma)$ before and after γ_b is that they are *nested*, as shown in figure 3.

Lemma 6. If there is no bifurcation point in the interval (σ_1, γ^+), then for any $\sigma_1 < \gamma'' < \gamma' < \gamma^+$ the two intersection curves $\mathbf{r}(\gamma')$ and $\mathbf{r}(\gamma'')$ are nested on the spherical surface, in the sense that (a) they do not intersect, and (b) the set of points on the unit sphere bounded by $\mathbf{r}(\gamma'')$ contains the set of points bounded by $\mathbf{r}(\gamma')$, including the initial tangency point $\mathbf{r}(\gamma^+)$. If, on the other hand, there is a bifurcation point in this interval, i.e., $\sigma_1 < \gamma_b$, then the statement above continues to hold for $\gamma_b < \gamma'' < \gamma' < \gamma^+$, while in the domain $\sigma_1 < \gamma'' < \gamma' < \gamma_b$ the containment is reversed: the set of points on the unit sphere bounded by $\mathbf{r}(\gamma'')$ is contained in the set of points bounded by any previous $\mathbf{r}(\gamma')$ including the limit curve $\mathbf{r}(\gamma_b)$.

Proof. Starting with two distinct intersection curves at $\gamma' \neq \gamma''$, one assumes that they intersect at $\mathbf{p} = \mathbf{r}(\gamma', \psi') = \mathbf{r}(\gamma'', \psi'')$. By subtracting the quadratic equation $\mathbf{p}^\top \mathbf{S}(\gamma')\mathbf{p} = \gamma'$ from $\mathbf{p}^\top \mathbf{S}(\gamma'')\mathbf{p} = \gamma''$ and simplifying, one gets the expression

$$1 = \sum_{k=1}^{N} \big(\mathbf{w}_k^\top \mathbf{p}\big)^2 \left(\frac{2\sigma_k - \gamma' - \gamma''}{(\gamma' - \sigma_k)^2 (\gamma'' - \sigma_k)^2} \right), \tag{47}$$

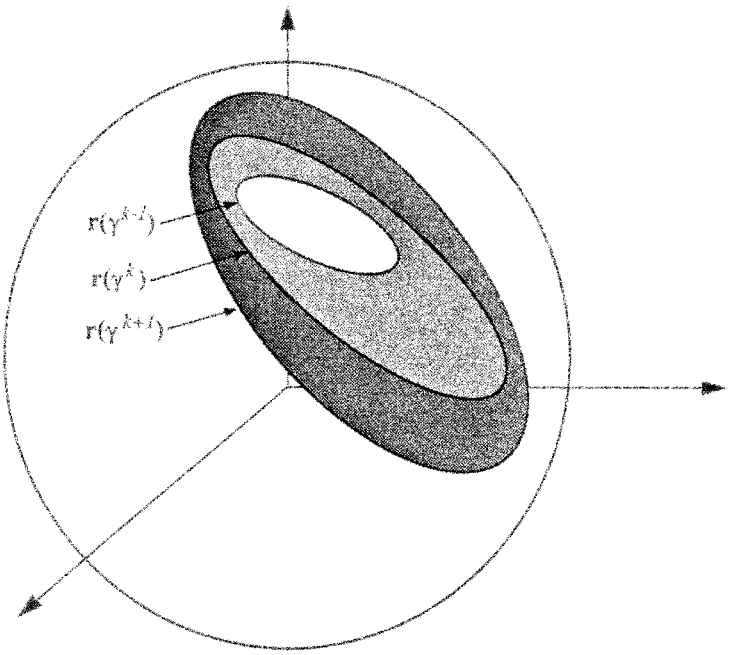

Figure 3. A nested sequence of intersection curves defined by a decreasing sequence of γ values.

whose right hand side is always negative for any choice of $\sigma_1 < \gamma'$, γ''. The contradiction in (47) can be traced back to the assumption that the curves intersect, which completes the proof on the nesting property. The two cases of containment are direct consequences of the inequalities among the projected semiaxes in equation (29), before and after the bifurcation point. □

Let $\chi(\gamma)$ be the maximum angle between any vector on the intersection curve $\mathbf{r}(\gamma)$ and its axis of symmetry. In view of lemma 4, this angle is given by

$$\chi(\gamma) = \sin^{-1}\big(\max(\beta_i, \beta_j)\big). \tag{48}$$

We shall now determine an upper bound γ_2^+ on the second eigenvalue by analyzing this angle. To start with, it can be identified with the half-angle of the circular cone having the same symmetry axis as $\mathbf{r}(\gamma)$ (see figure 4). Furthermore, the angle between any two points on the intersection curve $\mathbf{r}(\gamma)$ does not exceed $2\chi(\gamma)$. For that reason, we named γ_2^+ the "conical" bound. In what follows, we shall need the following definition:

$$\gamma_{b_1} = \begin{cases} \gamma_b & \text{if } \sigma_1 < \gamma_b, \\ \sigma_1 & \text{otherwise.} \end{cases} \tag{49}$$

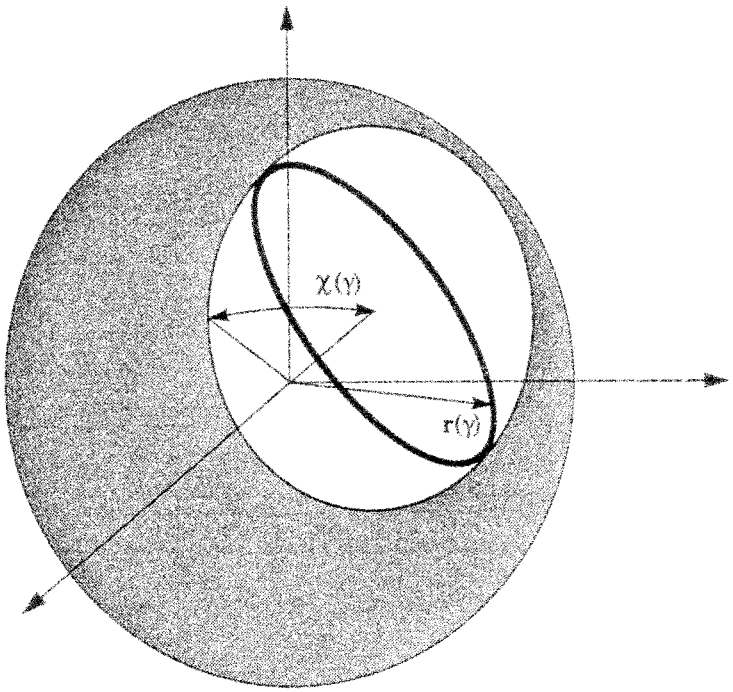

Figure 4. The cone defined by the intersection curve $\mathbf{r}(\gamma)$ and its half-angle $\chi(\gamma)$.

Theorem 7. The second eigenvalue is bounded from above by the unique solution γ_2^+ of the equation

$$\lambda_1(\gamma) + \lambda_2(\gamma) = 2 \tag{50}$$

in the interval (γ_{b_1}, γ^+).

Proof. By taking the sine of both sides of equation (48), followed by squaring, making use of equation (29) (recall that the eigenvalues of the matrix $\mathbf{S}(\gamma)/\gamma$ are ordered as $\lambda_3(\gamma) \leqslant \lambda_2(\gamma) \leqslant \lambda_1(\gamma)$) and rearranging, we find that equation (50) is equivalent to the condition $\chi(\gamma) = \pi/4$.

We now show that $\chi(\gamma) \geqslant \pi/4$ is necessary for the orthogonality of the first two eigenvectors. By lemma 6, the function $\chi(\gamma)$ is monotonically decreasing such that at the tangency point $\chi(\gamma^+) = 0$, while at a bifurcation point $\chi(\gamma_b) = \pi/2$, which follows from equation (55) and the fact that $\max(\beta_2, \beta_3) = 1$ at $\lambda_2 = 1$ by equation (29). If there is no bifurcation point in the interval (σ_1, γ^+), then σ_1 acts like one because when $\gamma \to \sigma_1$ then $\beta_2(\gamma) \to 1$ and $\chi(\gamma) \to \pi/2$. Hence, there is a unique value $\gamma_2^+ \in (\gamma_{b_1}, \gamma^+)$ at which $\chi(\gamma_2^+) = \pi/4$.

To show that γ_2^+ is an upper bound on γ_2, suppose first that the maximal eigenvalue satisfies $\gamma_1 \leqslant \gamma_2^+$, which does happen, though infrequently. In that case we must also have $\gamma_2 \leqslant \gamma_2^+$, since $\gamma_2 \leqslant \gamma_1$ by the ordering of the eigenvalues. In the more frequent case when $\gamma_2^+ < \gamma_1$, assume that $\gamma_2^+ < \gamma_2 \leqslant \gamma_1$. Then by lemma 6, the set

of points within the intersection curve $\mathbf{r}(\gamma_1)$ is contained in the set of points within the intersection curve $\mathbf{r}(\gamma_2)$, known to satisfy $2\chi(\gamma_2) < \pi/2$. It follows that there is no vector on the intersection curve $\mathbf{r}(\gamma_1)$ and no vector on the intersection curve $\mathbf{r}(\gamma_2)$ that could be mutually orthogonal and thus satisfy orthogonality condition (21). This contradiction is due to the assumption that $\gamma_2^+ < \gamma_2$ and we conclude that γ_2^+ must be an upper bound on γ_2, as well as on γ_3 by the ordering of the eigenvalues. □

To solve the nonlinear equation (50) we use the discretization of γ_1 over the interval $(\max(\sigma_1, \lambda_{\max}(\mathbf{G_0})), \gamma_1^+)$ that will be described in section 6.3. Upon bracketing the solution, we refine an initial solution by means of Newton's method. The calculation of the requisite eigenvalue derivatives λ_1' and λ_2' is described later on (equation (71)). This iteration is guaranteed to converge rapidly due to the monotonicity and differentiability of the underlying function in this interval.

In summary, we have shown that $\gamma_1 \in (\max(\sigma_1, \lambda_{\max}(\mathbf{G_0})), \gamma^+)$ and $\gamma_2, \gamma_3 \in (\gamma^-, \gamma_2^+)$ (see equation (46)). We note that we have never yet found an example in which $\gamma_3 < \sigma_3$ at the global minimum of f. This leads to the following:

Conjecture 1. At the global minimum of f the smallest eigenvalue satisfies $\gamma_3 > \max(\sigma_3, \gamma^-)$.

The conjectural part is the case when $\gamma^- < \sigma_3$, because the other case follows easily from the definition of γ^- in equation (46). If this conjecture turns out to be true, we will be able to further reduce the range of values accessible to the moments of inertia and hence reduce the time needed to find the global minimum. In the absence of a proof, we make sure that we find the global minimum by starting the search from γ^-, and accept some degradation of performance due to this less stringent lower bound.

5. Iterative methods

5.1. The Newton–Kantorovitch method for all variables

The first method of solving $\Phi(\mathbf{Y}) \equiv \nabla f(\mathbf{Y}) = \mathbf{0}$ is based on equation (10). It is a matrix version of the Newton–Kantorovitch method for solving operator equations [22]. The method exhibits quadratic convergence, and we have observed very few examples in which it fails to converge, although it may of course converge to a saddle point, local minimum or maximum.

Given any $\mathbf{Y}_n \in \mathbb{R}^{N\times 3}$ and a perturbation matrix $\Delta \in \mathbb{R}^{N\times 3}$, we expand the function $\Phi(\mathbf{Y}_n + \Delta)$ as

$$\begin{aligned}\Phi(\mathbf{Y}_n + \Delta) &= [\mathbf{Y}_n + \Delta]\big[(\mathbf{Y}_n + \Delta)^\top(\mathbf{Y}_n + \Delta) + \mathbf{X}^\top\mathbf{X}\big] - \mathbf{C}[\mathbf{Y}_n + \Delta] - \mathbf{V} \\ &= \Phi(\mathbf{Y}_n) + \mathbf{D}_\Delta(\Phi) + \mathbf{O}\big(\|\Delta\|^2\big), \end{aligned} \tag{51}$$

where $\mathbf{O}(\|\Delta\|^2)$ denotes matrices whose norm depends at least quadratically on $\|\Delta\|$. The terms linear in Δ are collected in the matrix

$$\mathbf{D}_\Delta(\Phi) = \Delta\big[\mathbf{Y}_n^\top \mathbf{Y}_n + \mathbf{X}^\top \mathbf{X}\big] + \mathbf{Y}_n \Delta^\top \mathbf{Y}_n + \big[\mathbf{Y}_n \mathbf{Y}_n^\top - \mathbf{C}\big]\Delta, \tag{52}$$

which is the directional or Frechet derivative of the matrix-valued function Φ evaluated in the direction Δ [22]. Setting the right-hand side of equation (51) to the zero matrix and ignoring the higher-order terms, we obtain a system of $3N$ linear equations in the $3N$ unknown entries of Δ, namely,

$$\mathbf{D}_\Delta(\Phi) = -\Phi(\mathbf{Y}_n), \tag{53}$$

whose solution Δ_n is used to update the next iterate as $\mathbf{Y}_{n+1} = \mathbf{Y}_n + \Delta_n$.

The perturbation matrix Δ appears linearly in equation (52), as does $\Delta^\top$, and each is multiplied on both sides by different matrix coefficients. A system of equations of this form is known as a *general linear system*. It is also known that in order to solve it by standard LU decomposition methods it must be converted into the form

$$\mathbf{K}_n \operatorname{col}(\Delta) = -\operatorname{col}\big(\Phi(\mathbf{Y}_n)\big). \tag{54}$$

Here, $\operatorname{col}(\cdot)$ is an operator that stacks the columns of its argument matrix into a single column vector. The construction of the matrix $\mathbf{K}_n \in \mathbb{R}^{3N\times 3N}$ is based on the Kronecker product $\otimes$, and the relation $\operatorname{col}(\mathbf{LMN}^\top) = (\mathbf{N}\otimes\mathbf{L})\operatorname{col}(\mathbf{M})$, which holds for arbitrary conformable matrices **L**, **M** and **N** [14]. On applying the col operator to equation (52) and using this relation, we find that

$$\begin{aligned}\operatorname{col}\big(\mathbf{D}_\Delta(\Phi)\big) = {} & \big[\big(\mathbf{Y}_n^\top \mathbf{Y}_n + \mathbf{X}^\top \mathbf{X}\big)\otimes \mathbf{I}_N\big]\operatorname{col}(\Delta) \\ & + \big[\mathbf{Y}_n^\top \otimes \mathbf{Y}_n\big]\operatorname{col}\big(\Delta^\top\big) + \big(\big[\mathbf{I}_3 \otimes \big(\mathbf{Y}_n \mathbf{Y}_n^\top - \mathbf{C}\big)\big]\operatorname{col}(\Delta)\big).\end{aligned} \tag{55}$$

To write this as a single linear system in $\operatorname{col}(\Delta)$, let $\mathbf{E}_{ij} \in \mathbb{R}^{N\times 3}$ be an elementary matrix, which has a one in the ijth position and zeros elsewhere, and let $\mathbf{P} \in \mathbb{R}^{3N\times 3N}$ be a permutation matrix whose columns are

$$\mathbf{P} = \big[\operatorname{col}(\mathbf{E}_{11}), \ldots, \operatorname{col}(\mathbf{E}_{13}), \operatorname{col}(\mathbf{E}_{21}), \ldots, \operatorname{col}(\mathbf{E}_{N3})\big].$$

It follows that $\operatorname{col}(\Delta^\top) = \mathbf{P}\operatorname{col}(\Delta)$ (see, e.g., [14]), and hence $\mathbf{K}_n$ is given by

$$\mathbf{K}_n = \big[\mathbf{Y}_n^\top \mathbf{Y}_n + \mathbf{X}^\top \mathbf{X}\big]\otimes \mathbf{I}_N + \big[\mathbf{Y}_n^\top \otimes \mathbf{Y}_n\big]\mathbf{P} + \mathbf{I}_3 \otimes \big[\mathbf{Y}_n \mathbf{Y}_n^\top - \mathbf{C}\big]. \tag{56}$$

Once $\operatorname{col}(\Delta_n) = -\mathbf{K}_n^{-1}\operatorname{col}(\Phi(\mathbf{Y}_n))$ has been obtained by standard numerical methods, it is easily put back into the desired increment matrix Δ_n.

The convergence of this iteration would be greatly improved if there were an inexpensive method of generating good starting matrices $\mathbf{Y}_0$. If $\|\mathbf{G_0}\|_1 \ll \gamma^-$ (see equation (46)), then we can assume $\mathbf{B} = \mathbf{0}$ to a good approximation, in which case the γ_k will be close to the three largest eigenvalues of **C** and we can use $\mathbf{Y}_0 = \widehat{\mathbf{Y}}\Gamma^{1/2}$, where the columns of $\widehat{\mathbf{Y}}$ are the corresponding eigenvectors. If $\|\mathbf{C}\|_1 \ll \lambda_{\min}(\mathbf{G_0})$, then we can simply set $\mathbf{Y}_0 = \mathbf{V}\mathbf{G_0}^{-1}$. More general procedures for finding good starting matrices, and searching the space for the global optimum, will be presented

later on. The Newton–Kantorovitch method has the advantage of being robust and easy to implement; it has the disadvantage of being relatively inefficient since it requires the repeated solution of a $3N \times 3N$ system of linear equations.

5.2. *Cayley parametrization of the inertial equations*

In this section we construct a method for solving equations (20)–(21) for the eigenvalues γ_1, γ_2, γ_3 and eigenvectors $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$ of the inertial tensor. This method is based on Cayley representation of an orthogonal matrix [1], which is possibly the simplest way to parametrize a rotation. Consider a real vector ω and the associated 3×3 skew-symmetric matrix:

$$\Omega = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}. \tag{57}$$

The *Cayley transform* is a rational matrix function which maps Ω into an orthogonal matrix $\mathbf{R}$:

$$\mathbf{R} = (\mathbf{I} - \Omega)/(\mathbf{I} + \Omega). \tag{58}$$

If $\|\omega\| = \sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2}$, the eigenvalues of $\mathbf{R}$ are given by

$$\left\{1, \frac{1 + \iota\|\omega\|}{1 - \iota\|\omega\|}, \frac{1 - \iota\|\omega\|}{1 + \iota\|\omega\|}\right\},$$

which shows that $\mathbf{R}$ is a proper orthogonal matrix, or rotation. The mapping $\omega \to \mathbf{R}$ is a bijection whose inverse is given by

$$\Omega = (\mathbf{I} - \mathbf{R})/(\mathbf{I} + \mathbf{R}) \tag{59}$$

whenever $\mathbf{R}$ is a rotation. By using the Cayley parametrization, we (a) eliminate all six constraints in equation (21) and hence reduce the total number of variables from twelve to six, and (b) have an efficient way of computing the vector ω for any given rotation matrix (cf. equation (59)).

In constructing $\mathbf{R}$ from the three eigenvectors $\mathbf{r}_1$, $\mathbf{r}_2$, $\mathbf{r}_3$, their arbitrary orientation causes an ambiguity in the values of ω. There are eight possible orientations, but four of these are eliminated since they result in $\det(\mathbf{R}) = -1$. Among the remaining four, we shall always choose the one that maximizes $\mathrm{tr}(\mathbf{R})$, because (a) it results in the smallest magnitude of $\|\omega\|$, and (b) improves the condition number of the matrix division in equation (59).

Appending the three rotational parameters ω_1, ω_2, ω_3 to the three eigenvalues γ_1, γ_2, γ_3, we form a vector of parameters $\mathbf{p} \in \mathbb{R}^6$, and then write the system of equations (20) in vector form

$$\mathbf{g}(\mathbf{p}) = \left[g_{11}(\mathbf{p})g_{22}(\mathbf{p})g_{33}(\mathbf{p})g_{12}(\mathbf{p})g_{13}(\mathbf{p})g_{23}(\mathbf{p})\right]^\top = \mathbf{0}. \tag{60}$$

Starting from a parameter vector $\mathbf{p}_0$, Newton's method for the system of equations (60) is given by $\mathbf{p}_{n+1} = \mathbf{p}_n - \mathbf{J}_n^{-1} \cdot \mathbf{g}_n$, where $\mathbf{g}_n = \mathbf{g}(\mathbf{p}_n)$ is the function vector and $\mathbf{J}_n = \partial\mathbf{g}/\partial\mathbf{p}|_{\mathbf{p}_n}$ is the Jacobian matrix evaluated at $\mathbf{p}_n$.

The partial derivatives of equation (60) with respect to the Cayley parameters ω_1, ω_2, ω_3 are computed as follows. Let $\Delta_m = \partial\Omega/\partial\omega_m$, $m = 1, 2, 3$, be the three directional derivative matrices of Ω, i.e.,

$$\Delta_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \Delta_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \qquad \Delta_3 = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{61}$$

Upon differentiating the expression $(\mathbf{I} + \Omega)(\mathbf{I} + \Omega)^{-1} = \mathbf{I}$ with respect to ω_m, we get

$$\Delta_m(\mathbf{I} + \Omega)^{-1} + (\mathbf{I} + \Omega)\frac{\partial}{\partial\omega_m}(\mathbf{I} + \Omega)^{-1} = \mathbf{0},$$

or

$$\frac{\partial}{\partial\omega_m}(\mathbf{I} + \Omega)^{-1} = -(\mathbf{I} + \Omega)^{-1}\Delta_m(\mathbf{I} + \Omega)^{-1}. \tag{62}$$

We now substitute equation (62) into

$$\frac{\partial\mathbf{R}}{\partial\omega_m} = (\mathbf{I} - \Omega)\frac{\partial}{\partial\omega_m}(\mathbf{I} + \Omega)^{-1} - \Delta_m(\mathbf{I} + \Omega)^{-1},$$

simplify it and obtain

$$\frac{\partial\mathbf{R}}{\partial\omega_m} = -2(\mathbf{I} + \Omega)^{-1}\Delta_m(\mathbf{I} + \Omega)^{-1}. \tag{63}$$

The three directional derivative matrices, $\partial\mathbf{R}/\partial\omega_m$, contain all nine derivative vectors $\partial\mathbf{r}_k/\partial\omega_m$ ($1 \leqslant k, m \leqslant 3$) and, as the columns of a rotation matrix, they are considered independent of the γ_k variables.

Since the vector $\mathbf{r}_k$ lies on the intersection curve $\mathbf{r}(\gamma_k)$ by lemma 3, one might ask why we ignore dependence of $\mathbf{r}_k$ on γ_k? The answer is that when we determine three unit vectors during the global search that approximately satisfy the nonlinear system of equations (60), they are only approximately orthogonal. Therefore, the inverse Cayley transform (equation (59)) results in an approximately skew-symmetric matrix $\widetilde{\Omega}$. When we skew-symmetrize it, according to $\overline{\Omega} = (\widetilde{\Omega} - \widetilde{\Omega}^{\top})/2$, in order to generate the three correct rotation parameters from which to start Newton's method, the corresponding rotation matrix $\overline{\mathbf{R}}$ has columns that do not lie any longer on the original intersection curves (generated by the γ_k) but are only in their vicinities. The result is that the terms containing $\mathrm{d}\mathbf{r}(\gamma_k)/\mathrm{d}\gamma_k|_{\mathbf{r}_k}$ can not be accurately calculated and hence they are omitted from equations (64) below. The price for this simplification of Jacobian matrix is that the convergence rate is somewhat less than quadratic.

Therefore the formulae for the entries of the Jacobian matrix take on a relatively simple form, namely,

$$\begin{aligned}\frac{\partial g_{kk}}{\partial \gamma_k} &= 1 + 2\mathbf{r}_k^\top \big[\mathbf{G_0} + \mathbf{W}^\top \big([\gamma_k \mathbf{I} - \Sigma]^{-3}\big)\mathbf{W}\big]\mathbf{r}_k, \qquad 1 \leqslant k \leqslant 3,\\ \frac{\partial g_{kl}}{\partial \gamma_k} &= -\mathbf{r}_k^\top \big[\mathbf{G_0} + \mathbf{W}^\top \big([\gamma_k \mathbf{I} - \Sigma]^{-2}[\gamma_l \mathbf{I} - \Sigma]^{-1}\big)\mathbf{W}\big]\mathbf{r}_l, \quad 1 \leqslant k < l \leqslant 3.\end{aligned} \tag{64}$$

Finally, by equations (18)–(19), $\mathbf{S}(\gamma_k, \gamma_l) \to \mathbf{S}(\gamma_k)$ as $\gamma_l \to \gamma_k$, so that

$$\frac{\partial g_{kl}}{\partial \omega_m} = \left[\frac{\partial \mathbf{r}_k}{\partial \omega_m}\right]^\top \mathbf{S}(\gamma_k, \gamma_l)\mathbf{r}_l + \left[\frac{\partial \mathbf{r}_l}{\partial \omega_m}\right]^\top \mathbf{S}(\gamma_k, \gamma_l)\mathbf{r}_k, \quad 1 \leqslant k \leqslant l \leqslant 3. \tag{65}$$

The rest of the entries are equal to zero.

5.3. *Angular parametrization of the inertial equations*

An alternative parametrization which reduces equations (20)–(21) to a system of six equations in six unknowns uses the three eigenvalues γ_k together with the three angles ψ_k defined in equation (30), one for each intersection curve $\mathbf{r}(\gamma_k)$. For any three values of γ_k (satisfying the intersection condition (27)) and any three values of $\psi_k \in [0, 2\pi)$, the three unit vectors $\mathbf{r}(\gamma_k, \psi_k)$ satisfy the three quadratic equations (20) and the three normalizing conditions in equations (21), but are not necessarily orthogonal. Conversely, for each column vector $\mathbf{r}_k$ on the intersection curve $\mathbf{r}(\gamma_k)$ the corresponding angular parameter ψ_k is uniquely determined by equations (29)–(31).

If we denote the parameter vector $\mathbf{q} = [\gamma_1\ \psi_1\ \gamma_2\ \psi_2\ \gamma_3\ \psi_3]^\top$, then the remaining three equations (20) can be written as

$$\begin{aligned} h_1(\gamma_1, \psi_1, \gamma_2, \psi_2) &\equiv \mathbf{r}^\top(\gamma_1, \psi_1)\mathbf{S}(\gamma_1, \gamma_2)\mathbf{r}(\gamma_2, \psi_2) = 0,\\ h_3(\gamma_2, \psi_2, \gamma_3, \psi_3) &\equiv \mathbf{r}^\top(\gamma_2, \psi_2)\mathbf{S}(\gamma_2, \gamma_3)\mathbf{r}(\gamma_3, \psi_3) = 0,\\ h_5(\gamma_3, \psi_3, \gamma_1, \psi_1) &\equiv \mathbf{r}^\top(\gamma_1, \psi_1)\mathbf{S}(\gamma_1, \gamma_3)\mathbf{r}(\gamma_3, \psi_3) = 0,\end{aligned} \tag{66}$$

while the three orthogonality conditions in equations (21) become

$$\begin{aligned} h_2(\gamma_1, \psi_1, \gamma_2, \psi_2) &\equiv \mathbf{r}^\top(\gamma_1, \psi_1)\mathbf{r}(\gamma_2, \psi_2) = 0,\\ h_4(\gamma_2, \psi_2, \gamma_3, \psi_3) &\equiv \mathbf{r}^\top(\gamma_2, \psi_2)\mathbf{r}(\gamma_3, \psi_3) = 0,\\ h_6(\gamma_3, \psi_3, \gamma_1, \psi_1) &\equiv \mathbf{r}^\top(\gamma_1, \psi_1)\mathbf{r}(\gamma_3, \psi_3) = 0.\end{aligned} \tag{67}$$

This completes the description of the system of equations.

We now illustrate the computation of the entries of the Jacobian matrix by presenting only the partial derivatives for the functions h_1 and h_2 with respect to γ_1

and ψ_1. The remaining derivatives have the same form apart from index permutation. The partial derivatives with respect to ψ_1 are straightforward:

$$\begin{aligned} \frac{\partial h_1}{\partial \psi_1} &= \frac{\partial \mathbf{r}^\top(\gamma_1,\psi_1)}{\partial \psi_1}\,\mathbf{S}(\gamma_1,\gamma_2)\mathbf{r}(\gamma_2,\psi_2), \\ \frac{\partial h_2}{\partial \psi_1} &= \frac{\partial \mathbf{r}^\top(\gamma_1,\psi_1)}{\partial \psi_1}\,\mathbf{r}(\gamma_2,\psi_2), \end{aligned} \tag{68}$$

because $\mathbf{r}(\gamma_2,\psi_2)$ does not depend on ψ_1. The partial derivative of the curve $\mathbf{r}(\gamma_1,\psi_1)$ with respect to ψ_1 is obtained via equations (29)–(31) as

$$\frac{\partial \mathbf{r}(\gamma_1,\psi_1)}{\partial \psi_1} = \mathbf{U}(\gamma_1)\begin{bmatrix} \partial\xi_1/\partial\psi_1 \\ \partial\xi_2/\partial\psi_1 \\ \partial\xi_3/\partial\psi_1 \end{bmatrix} = \mathbf{U}(\gamma_1)\begin{bmatrix} -\beta_1\sin(\psi_1) \\ \beta_2\cos(\psi_1) \\ \dfrac{(\beta_2-\beta_1)\cos(\psi_1)\sin(\psi_1)}{\xi_3} \end{bmatrix} \tag{69}$$

if, for example, the axis of symmetry is $\mathbf{u}_3$.

The partial derivatives of h_1 and h_2 with respect to the eigenvalue γ_1 are more complicated. They depend on the derivatives of the spectral decomposition of the matrix $\mathbf{Q}(\gamma) = \mathbf{S}(\gamma)/\gamma$ (see section 4.1). The derivatives of the eigenvalues and eigenvectors of a matrix-valued function of a scalar argument have been rediscovered many times, see, e.g., [2,19] and references therein. In physics, the eigenvalue derivatives are implicit in the Hellmann–Feynman theorem [6], but the eigenvector derivatives appear to be much less widely known. For this reason we include a brief derivation of these derivatives for the case of a symmetric (or hermitian) matrix.

Assuming that $\mathbf{Q}(\gamma) = \mathbf{U}(\gamma)\Lambda(\gamma)\mathbf{U}^\top(\gamma)$ is a continuously differentiable symmetric matrix function in some domain, its derivative with respect to γ is

$$\mathbf{Q}' = \mathbf{U}'\Lambda\mathbf{U}^\top + \mathbf{U}\Lambda'\mathbf{U}^\top + \mathbf{U}\Lambda\mathbf{U}'^\top.$$

The relation $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ implies $(\mathbf{U}^\top\mathbf{U}')^\top = -\mathbf{U}^\top\mathbf{U}'$ and hence

$$\Lambda' + [\Lambda, \mathbf{M}] = \mathbf{U}^\top\mathbf{Q}'\mathbf{U}, \tag{70}$$

where the square brackets denote the matrix commutator, and the matrix $\mathbf{M} = -\mathbf{U}^\top\mathbf{U}'$ is variously known as the Cartan matrix, logarithmic derivative, or multiplicative derivative of $\mathbf{U}^\top$ [12]. Since the commutator of any matrix with a diagonal one has zeros on its diagonal, it follows that

$$\Lambda' = \operatorname{diag}\bigl(\mathbf{U}^\top\mathbf{Q}'\mathbf{U}\bigr). \tag{71}$$

The matrix $\mathbf{U}^\top\mathbf{Q}'\mathbf{U}$ is symmetric (though not diagonal, because $[\mathbf{Q},\mathbf{Q}'] \neq \mathbf{0}$), which implies that the commutator $[\Lambda,\mathbf{M}]$ is symmetric too. Thus the entries m_{ij} are obtained

by simply writing out the off-diagonal entries in equation (70) and then solving the resulting (scalar) linear equations, namely,

$$m_{ij} = -\mathbf{u}_i^\top \mathbf{u}'_j = \frac{\mathbf{u}_i^\top \mathbf{Q}' \mathbf{u}_j}{\lambda_i - \lambda_j} \quad (i \neq j). \tag{72}$$

Finally, the eigenvector derivatives are computed from $\mathbf{U}' = -\mathbf{U}\mathbf{M}$. When $\lambda_i = \lambda_j$ for a particular value of γ, a further analysis is required which we shall not pursue, because this case has not occurred in the applications of interest here.

Given the eigenvalue and eigenvector derivatives of $\mathbf{Q}(\gamma)$, the functions h_1 and h_2 are differentiated with respect to γ_1 as follows:

$$\begin{aligned} \frac{\partial h_1}{\partial \gamma_1} &= \left(\frac{\partial \mathbf{r}^\top(\gamma_1, \psi_1)}{\partial \gamma_1} \mathbf{S}(\gamma_1, \gamma_2) + \mathbf{r}^\top(\gamma_1, \psi_1) \frac{\partial \mathbf{S}(\gamma_1, \gamma_2)}{\partial \gamma_1} \right) \mathbf{r}(\gamma_2, \psi_2), \\ \frac{\partial h_2}{\partial \gamma_1} &= \frac{\partial \mathbf{r}^\top(\gamma_1, \psi_1)}{\partial \gamma_1} \mathbf{r}(\gamma_2, \psi_2). \end{aligned} \tag{73}$$

The matrix derivative in equation (73) is simply

$$\frac{\partial \mathbf{S}(\gamma_1, \gamma_2)}{\partial \gamma_1} = -\mathbf{W}^\top \left(\left[\gamma_1 \mathbf{I}_N - \mathbf{\Sigma} \right]^{-2} \left[\gamma_2 \mathbf{I}_N - \mathbf{\Sigma} \right]^{-1} \right) \mathbf{W}, \tag{74}$$

while the partial derivative of the intersection curve with respect to γ_1,

$$\frac{\partial \mathbf{r}(\gamma_1, \psi_1)}{\partial \gamma_1} = \frac{\mathrm{d}\mathbf{U}(\gamma_1)}{\mathrm{d}\gamma_1} \xi(\gamma_1, \psi_1) + \mathbf{U}(\gamma_1) \frac{\partial \xi(\gamma_1, \psi_1)}{\partial \gamma_1}, \tag{75}$$

requires the matrix $\mathrm{d}\mathbf{U}(\gamma_1)/\mathrm{d}\gamma_1$ (obtained via equation (72)), and the partial derivative vector $\partial \xi / \partial \gamma_1$. The latter depends on the ordinary derivatives of the projected semiaxes (see equation (29)), for example,

$$\frac{\mathrm{d}\beta_i}{\mathrm{d}\gamma_1} = \begin{cases} \dfrac{(1-\lambda_2)\lambda'_1 + (\lambda_1 - 1)\lambda'_2}{2\beta_i(\lambda_1 - \lambda_2)^2} & \text{if } \lambda_2 < 1, \\ \dfrac{(1-\lambda_3)\lambda'_1 + (\lambda_1 - 1)\lambda'_3}{2\beta_i(\lambda_1 - \lambda_3)^2} & \text{if } \lambda_2 > 1, \end{cases} \tag{76}$$

which in turn depend on the eigenvalue derivatives $\lambda'_m = \mathrm{d}\lambda_m/\mathrm{d}\gamma_1$ (see equation (71)). Thus, the expression for $\partial \xi / \partial \gamma_1$ is derived by using equations (29), (30), (71) and (76).

6. The search for the global minimum

6.1. An overview of the search strategy

All three variations on Newton's method in section 5 may diverge if they are not started sufficiently close to a solution. In this regard, the Newton–Kantorovitch method

tends to be superior to the other two, but it requires the solution of a $3N \times 3N$ system of linear equations at each iteration, whereas the two based on the inertial equations require the solution of only a 6×6 system. Furthermore, due to the existence of several solutions, none of these methods can be expected to find, on its own and with any degree of reliability, the *global* minimum of the *STRAIN* f, which is what interests us the most.

For this reason we have developed a search procedure that is capable of finding all solutions to the inertial equations in the subspace defined by the bounds established in section 4. The global minimum is then easily found by constructing the associated critical matrices via equation (25), evaluating the function f at each, and selecting the minimal one. Numerical evidence gathered over numerous problems of size $M + N < 200$ indicates that the number of such critical matrices rarely exceeds a dozen or so, and hence this search is generally quite efficient. The procedure as a whole consists of four stages.

In the first stage, the two γ-intervals defined by the bounds

$$\gamma_1 \in \big(\max\big(\sigma_1, \lambda_{\max}(\mathbf{G_0})\big), \gamma^+\big)$$

and $\gamma_2, \gamma_3 \in (\gamma^-, \gamma_2^+)$ are discretized subject to a given curve length increment criterion with a deliberately increased discretization density in the vicinity of poles and bifurcation points. The outcome is two nonuniform discrete sequences, or meshes, one for γ_1, the other shared by both γ_2, γ_3. This stage of the algorithm takes a small fraction of the overall time.

The main idea comes from lemma 3, which states that each discrete value of γ_1 in its range defines a unique intersection curve $\mathbf{r}(\gamma_1)$ which contains *all* solutions $\mathbf{r}$ to the first inertial equation $\gamma_1 - \mathbf{r}^\top \mathbf{S}(\gamma_1)\mathbf{r} = 0$. By discretizing this curve, as described in section 6.2, we generate a discrete set of all inertial pairs $\{\gamma_1, \mathbf{r}_1\}$ at fixed γ_1. We shall do that for each curve defined by the γ_1-mesh and so obtain a sample covering the portion of the unit sphere of interest. The discretization process for the γ-intervals is described in section 6.3 below.

The second stage systematically generates a discrete set of trial solutions over the γ-meshes, where each trial solution consists of three distinct inertial pairs. Each trial solution is tested to see if it satisfies the three bilinear inertial equations (20) to within a given threshold, and if so, it is appended to a list of initial solutions. These generation and elimination processes are described in section 6.4. This stage of the algorithm may take a significant portion of the overall time, and is problem dependent.

The third and the fourth stage operate together as follows: each initial solution from the list is refined by means of one of the three versions of Newton's method derived in section 5. In the case of convergence, the function f is evaluated at this new critical matrix and inserted into an ordered list of the function values already found, provided it is different from them. These two final stages of the algorithm are proportional to the number of initial points generated in the second stage, and may also take significant portion of the overall time.

6.2. Discretization of the intersection curves

For any given γ_1 in the mesh over the interval in equation (35), let $\mathbf{r}(\gamma_1, \psi_1)$ be an intersection curve parametrized by ψ_1. We choose as our criterion for discretizing it a fixed spherical distance $\Delta\theta$ between successive points along the curve, namely,

$$\mathbf{r}^\top(\gamma_1, \psi_1^k)\mathbf{r}(\gamma_1, \psi_1^{k+1}) = \cos(\Delta\theta). \tag{77}$$

Since we will be dealing in this section with a single fixed value of γ_1, we shall omit the subscript from ψ_1^k and the dependence on γ_1 altogether, denoting the points on the intersection curve simply by $\mathbf{r}(\psi^k)$. The problem then is to find a discrete sequence of parameters $\{\psi^k\}_{k=1}^m$ satisfying equation (77).

Due to the double symmetry of the intersection curve (see section 4.1), it is sufficient to consider only the interval $[0, \pi/2]$, and find a subsequence of parameters $0 = \psi^0 < \psi^1 < \cdots < \psi^{m/4} = \pi/2$ satisfying (77). The complete sequence is then obtained by reflecting and copying this subsequence to obtain exactly m points, as follows:

$$\begin{aligned}\big\{&\psi^0, \ldots, \psi^{m/4}, \pi - \psi^{m/4-1}, \ldots, \pi - \psi^0,\\ &\pi + \psi^1, \ldots, \pi + \psi^{m/4}, 2\pi - \psi^{m/4-1}, \ldots, 2\pi - \psi^1\big\}.\end{aligned} \tag{78}$$

Suppose we have found points $\psi^0, \ldots, \psi^k$ satisfying equation (77), and such that $\psi^k < \pi/2$. Finding the next point means solving the equation

$$\mathbf{r}^\top(\psi^k)\mathbf{r}(\psi^k + \delta\psi) = \cos(\Delta\theta) \tag{79}$$

for the increment $\delta\psi$. While it is certainly possible to solve equation (70) at each step to full working precision, the final parameter will in general exceed $\pi/2$. If the last parameter is reset to $\pi/2$, then the last spherical distance may significantly differ from others. A slight redistribution of the points, on the other hand, would accommodate for this nonuniformity but then some of the work in getting the exact solution would have been wasted. A way out of this situation comes from the following argument. Consider the derivative of the arc length of the intersection curve with respect to the parameter ψ, i.e.,

$$s'(\psi) = \sqrt{\frac{\beta_i^2 \sin^2(\psi) - \beta_j^2(\beta_i^2 - \cos^2(\psi))}{1 - \beta_i^2\cos^2(\psi) - \beta_j^2\sin^2(\psi)}}, \tag{80}$$

which is obtained from $s'(\psi) = \sqrt{(\xi_1'(\psi))^2 + (\xi_2'(\psi))^2 + (\xi_3'(\psi))^2}$ and the derivatives of equation (30), where $0 < \beta_i, \beta_j < 1$ are the semi-axes of the projected ellipse. A quick analysis shows that $s'(\psi)$ is a periodic function with period π, which is monotone on the intervals $[0, \pi/2]$ and $[\pi/2, \pi]$, but with different monotonicity on latter due to the reflection $s'(\psi) = s'(\pi - \psi)$ in $\pi/2$. Thus, $s'(\psi)$ attains its extremal values $\min(\beta_i, \beta_j)$ and $\max(\beta_i, \beta_j)$ at the points 0 and $\pi/2$. Since the arc length between any two points

on a continuous spherical curve is bounded from below by their spherical distance, we see that

$$\begin{aligned}\Delta\theta &= \operatorname{acos}\left(\mathbf{r}^\top(\psi^k)\mathbf{r}(\psi^k+\delta\psi)\right) \leqslant \delta s^k \\ &= \int_{\psi^k}^{\psi^k+\delta\psi} s'(\psi)\,\mathrm{d}\psi \leqslant \max(\beta_i,\beta_j)\delta\psi. \end{aligned} \tag{81}$$

It follows that $\Delta\theta/\max(\beta_i,\beta_j)$ is a lower bound on all the parametric increments $\delta\psi$ satisfying equation (79). Locally, however, a better approximation is given by $\delta\psi \approx \Delta\theta/s'(\psi^k)$. The following discretization algorithm is based on this idea:

$$\begin{array}{ll} \texttt{set:} & \psi^0 = 0,\ k = 0 \\ \texttt{while} & (\psi^k < \pi/2) \\ & \quad \delta\psi^k = \Delta\theta/s'(\psi^k) \\ & \quad \psi^{k+1} = \psi^k + \delta\psi^k \\ & \quad k = k+1 \\ \texttt{end} & \end{array} \tag{82}$$

Upon exiting the while loop, we set $m = 4k$ and then correct for any "overshoot" $\psi^{m/4} > \pi/2$ by rescaling the sequence of increments $\delta\psi^k$ using the relative scale $\eta = \pi/(2\psi^{m/4})$. This means that a new sequence of angles is generated by $\psi^k = \psi^{k-1} + \delta\psi^{k-1}\eta$ for $k = 1,\ldots,m/4$.

This algorithm produces sequences that are approximately uniform according to the criterion in equation (79). In addition, we make sure that it slightly oversamples the interval by starting the sequence from the point with the maximum derivative of the arc length, i.e., $\psi^0 = 0$ if $\beta_i \leqslant \beta_j$ or $\psi^0 = \pi/2$ if $\beta_i \geqslant \beta_j$, since then $s'(\psi^k) > s'(\psi^{k+1})$. In the last case, we first subtract $\pi/2$ from the computed subsequence before reflecting and copying as in equation (78).

6.3. *Discretization of the γ intervals*

In order to discretize the two γ intervals, $(\max(\sigma_1, \lambda_{\max}(\mathbf{G_0})), \gamma^+)$ and (γ^-, γ_2^+), we need a suitable definition of the distance between two consecutive nested curves $\mathbf{r}(\gamma^i)$ and $\mathbf{r}(\gamma^{i+1})$. We have considered a number of such definitions, e.g., the Hausdorff distance and several of its approximations, or the half-angle $\chi(\gamma)$ of the enveloping cone, but they resulted in either too many points, or are not sufficiently sensitive in the regions of fast variation of an intersection curve such as the tangency and bifurcation points.

It is these considerations that led us to consider the length of an intersection curve $0 \leqslant L(\gamma) \leqslant 2\pi$ as a quantity which is continuous, and inherits the behavior of both $\beta_i(\gamma)$ and $\beta_j(\gamma)$. Furthermore, if there is a bifurcation point $\gamma_b \in (\max(\sigma_1, \lambda_{\max}(\mathbf{G_0})), \gamma^+)$, then $L(\gamma)$ is a monotonically increasing function on the interval $\gamma \in (\max(\sigma_1, \lambda_{\max}(\mathbf{G_0})), \gamma_b)$, and a monotonically decreasing one on the interval $\gamma \in (\gamma_b, \gamma^+)$, with its maximum attained at the bifurcation point. In

spite of the jump discontinuity in the values of $\beta_i(\gamma)$ and $\beta_j(\gamma)$ across γ_b, when we evaluate arc length derivatives in equation (80) on both sides of γ_b, we find that $s'(\gamma_b^-, \psi) = s'(\gamma_b^+, \psi) = 1$ for all ψ, confirming that $\max_\gamma(L(\gamma)) = L(\gamma_b) = 2\pi$. The derivative $L'(\gamma)$ is also well defined, except at the tangency points where $L(\gamma_t) = 0$ and $L'(\gamma_t) = \pm\infty$ due to a square root type singularity. At a bifurcation point we have

$$\begin{aligned} \lim_{\gamma\to\gamma_b^+} \big(L'(\gamma)\big) &= 0, \\ \lim_{\gamma\to\gamma_b^-} \big(L'(\gamma)\big) &= 0. \end{aligned} \tag{83}$$

The length of an intersection curve is obtained by integrating the expression in equation (80) (which really is the partial derivative $s' = \partial s(\gamma, \psi)/\partial\psi$) and taking advantage of its double symmetry, i.e.,

$$L(\gamma) = 4\int_0^{\pi/2} \frac{\partial s(\gamma, \psi)}{\partial\psi}\,\mathrm{d}\psi. \tag{84}$$

For $\mathrm{d}L/\mathrm{d}\gamma$ we calculate the mixed partial derivative of $s(\gamma, \psi)$, so that

$$\frac{\mathrm{d}L}{\mathrm{d}\gamma} = 4\int_0^{\pi/2} \frac{\partial^2 s(\gamma, \psi)}{\partial\gamma\partial\psi}\,\mathrm{d}\psi. \tag{85}$$

The interchange of integration and partial differentiation is permissible for any region not containing the points γ_b, γ_t, σ_i. Both integrals are evaluated using standard Romberg quadrature and the analytic expressions for $\partial s(\gamma, \psi)/\partial\psi$ (see equation (80)) and $\partial^2 s(\gamma, \psi)/(\partial\gamma\,\partial\psi)$ (equations (71), (76) and (80)). The maximal number of interval halvings rarely exceeded seven for double precision accuracy.

In many ways the γ-discretization can be made similar to the ψ-discretization. This means that if we specify the length difference between two consecutive curves, namely $\Delta L = 2\pi/m$ for some $m > 1$, then instead of solving the nonlinear equation

$$L(\gamma^k + \delta\gamma) - L(\gamma^k) - \Delta L = 0 \tag{86}$$

exactly for the next increment $\delta\gamma$, we use an approximate version of equation (86), namely,

$$\gamma^{k+1} = \gamma^k + \frac{\Delta L}{L'(\gamma^k)}. \tag{87}$$

This will generate a decreasing sequence γ^k, providing it starts slightly to the left of γ^+ (to avoid division by $-\infty$), and ending slightly to the right of $\max(\sigma_1, \gamma_b)$ (to avoid division by zero). In the case that $\sigma_1 < \gamma_b$ we continue generating the sequence of γ^k using the same equation (87), but with negative ΔL since the lengths are decreasing. As before, we start it slightly to the left of γ_b, and end it slightly to the right from σ_1. A small interval around a bifurcation point is discretized more densely in order to capture small or ill-conditioned intersection curves that are associated with them.

As mentioned earlier, when a bifurcation point is close to a pole, it often generates in its vicinity very thin intersection curves that appear as very narrow cuts on the unit sphere. Two kinds of problems occur in the presence of such an ill conditioned geometry: (a) the search process is not able to locate relevant initial solutions, and (b) Newton's method tends to diverge more frequently. We therefore increase the density of the γ-mesh in an inversely proportional manner to the eccentricity $\min(\beta_i/\beta_j, \beta_j/\beta_i)$ in any region where this number drops below a given threshold. This mitigates the aforementioned difficulties.

6.4. Elimination criteria for the global search

We recall that for each $\gamma_1 \in (\max(\sigma_1, \lambda_{\max}(\mathbf{G_0})), \gamma^+)$ and for each $\psi_1 \in [0, 2\pi)$, the inertial pair $\gamma_1, \mathbf{r}_1 = \mathbf{r}(\gamma_1, \psi_1)$ satisfies the first inertial equation $\mathbf{r}_1^\top \mathbf{S}(\gamma_1)\mathbf{r}_1 = \gamma_1$. There are five more equations left to solve but only three variables, since we have fixed γ_1 and fixing the unit vector $\mathbf{r}_1$ eliminates *two* degrees of freedom. Two of these remaining variables are γ_2 and γ_3, while the third is chosen to be the angle ϕ in the plane orthogonal to $\mathbf{r}_1$ in which $\mathbf{r}_2$ and $\mathbf{r}_3$ must lie in order to form an orthonormal frame. Let $\mathbf{N} = [\mathbf{n}_1, \mathbf{n}_2]$ be any 3×2 matrix such that $[\mathbf{r}_1, \mathbf{n}_1, \mathbf{n}_2]$ forms a proper orthonormal frame. Then the vectors $\mathbf{r}_2$, $\mathbf{r}_3$ are expressed as

$$\begin{aligned} \mathbf{r}_2 &= \mathbf{n}_1 \cos(\phi) + \mathbf{n}_2 \sin(\phi) = \mathbf{N}\mathbf{t}_2, \\ \mathbf{r}_3 &= -\mathbf{n}_1 \sin(\phi) + \mathbf{n}_2 \cos(\phi) = \mathbf{N}\mathbf{t}_3, \end{aligned} \tag{88}$$

where $\mathbf{t}_2^\top = [\cos(\phi), \sigma(\phi)]$ and $\mathbf{t}_3^\top = [-\sin(\phi), \cos(\phi)]$ are evidently orthonormal plane vectors. We now rewrite the remaining two quadratic inertial equations in equations (20) in terms of this new angular variable ϕ as

$$\begin{aligned} \mathbf{t}_2^\top(\phi)\mathbf{N}^\top \mathbf{S}(\gamma_2)\mathbf{N}\mathbf{t}_2(\phi) &= \gamma_2, \\ \mathbf{t}_3^\top(\phi)\mathbf{N}^\top \mathbf{S}(\gamma_3)\mathbf{N}\mathbf{t}_3(\phi) &= \gamma_3. \end{aligned} \tag{89}$$

These two equations differ only by the subscripts on γ and $\mathbf{t}$, and thus the solutions $\{\gamma, \phi\}$ to one of them are also the solutions to the other (recall that γ_2, γ_3 share the same domain as well as mesh). The same observation holds for the two bilinear equations g_{12}, g_{13}: they are identical, apart from the same two subscripts, namely,

$$\begin{aligned} \mathbf{r}_1^\top \mathbf{S}(\gamma_1, \gamma_2)\mathbf{N}\mathbf{t}_2(\phi) &= 0, \\ \mathbf{r}_1^\top \mathbf{S}(\gamma_1, \gamma_3)\mathbf{N}\mathbf{t}_3(\phi) &= 0. \end{aligned} \tag{90}$$

These facts, which are a consequence of the formal symmetry of the system of equations (20), provide for considerable economy of effort. In the following analysis of these equations, therefore, we shall replace the subscript 2 or 3 with t whenever the same argument applies equally to both of the equations in (89) or (90).

The 2×2 matrix $\mathbf{N}^\top[\mathbf{S}(\gamma_t)/\gamma_t]\mathbf{N}$ in equation (89) determines an ellipse, which is obtained by intersecting the ellipsoid (defined by $\mathbf{S}(\gamma_t)/\gamma_t$) with the central plane spanned by $[\mathbf{n}_1, \mathbf{n}_2]$. Therefore the solutions of equation (89) are the intersections of

the intersection curves (between the ellipsoid and the unit sphere) with the plane **N**. This is clearly a triple surface intersection problem, made considerably easier by the central symmetry of all three surfaces. Such an intersection exists if and only if the eigenvalues $\mu_{\min}(\gamma_t)$ and $\mu_{\max}(\gamma_t)$ of $\mathbf{N}^\top[\mathbf{S}(\gamma_t)/\gamma_t]\mathbf{N}$ satisfy

$$\mu_{\min}(\gamma_t) \leqslant 1 \leqslant \mu_{\max}(\gamma_t). \tag{91}$$

Note also that, by the Rayleigh quotient argument, the inequalities

$$\lambda_3\big(\mathbf{S}(\gamma_t)/\gamma_t\big) \leqslant \mathbf{n}^\top\big[\mathbf{S}(\gamma_t)/\gamma_t\big]\mathbf{n} \leqslant \lambda_1\big(\mathbf{S}(\gamma_t)/\gamma_t\big)$$

hold for any unit vector **n**, and in particular for $\mathbf{n} = \mathbf{N}\mathbf{p}$, where **p** is any unit vector in the plane. This shows that $\lambda_3 \leqslant \mu_{\min}$ and $\mu_{\max} \leqslant \lambda_1$, so it is quite possible for the intersection curves not to intersect a given plane **N**, and hence it is always necessary to check the inequalities in equation (91) directly.

Let $\mathbf{P} = [\mathbf{p}_{\min}, \mathbf{p}_{\max}]$ be the eigenvectors of $\mathbf{N}^\top[\mathbf{S}(\gamma_t)/\gamma_t]\mathbf{N}$ corresponding to the eigenvalues $\mu_{\min}$, $\mu_{\max}$. When the intersection condition (91) holds, the two exact intersection points for equation (89) are given by

$$\mathbf{t}_\pm = \mathbf{P}\mathbf{m}_\pm, \qquad \mathbf{m}_\pm = \frac{1}{\sqrt{\mu_{\max} - \mu_{\min}}} \begin{bmatrix} \sqrt{1-\mu_{\min}} \\ \pm\sqrt{\mu_{\max}-1} \end{bmatrix}. \tag{92}$$

Their negatives, $-\mathbf{r}_\pm = -\mathbf{N}\mathbf{t}_\pm$, are also intersection points but, as we have seen, they are redundant in the sense that they give rise to the same critical matrices of f, and hence need not be considered.

The above arguments can be restated as follows: for any fixed inertial pair $\{\gamma_1, \mathbf{r}_1\}$ and any γ_t taken from its mesh, either there are no solutions (criterion (91) is not satisfied) or else two solutions $\mathbf{r}_\pm$ are calculated from equations (88) and (92). We now subject both solutions to the first test, based on the bilinear equation (90) and an orthogonality threshold $\varepsilon_\perp$, i.e.,

$$\big|\mathbf{r}_1^\top \mathbf{S}(\gamma_1, \gamma_t)\mathbf{r}_\pm\big| < \varepsilon_\perp \big\|\mathbf{S}(\gamma_1, \gamma_t)\big\|_1. \tag{93}$$

This means that if an inertial pair $\{\gamma_t, \mathbf{r}_t\}$ satisfies this inequality, then it is an approximate solution to $g_{12}(\gamma_1, \mathbf{r}_1, \gamma_t, \mathbf{r}_t) = 0$ as well as an exact solution to the quadratic equation $g_{tt}(\gamma_t, \mathbf{r}_t) = 0$. Each *inertial pair* $\{\gamma_t, \mathbf{r}_t\}$ that satisfies this test is appended as $\{\gamma^l, \mathbf{r}^l\}$ to a temporary list L_T (associated with a single $\{\gamma_1, \mathbf{r}_1\}$), which will be used in the next step of the search.

In this next step, two additional tests are applied to each *pair* of elements $\{\gamma^{l_1}, \mathbf{r}^{l_1}\}$, $\{\gamma^{l_2}, \mathbf{r}^{l_2}\}$ from the list L_T. First, we check that the pair of vectors is approximately orthogonal, i.e.,

$$\big(\mathbf{r}^{l_1}\big)^\top \mathbf{r}^{l_2} < \varepsilon_\perp. \tag{94}$$

Second, we check that each pair that passes test (94) also approximately satisfies the last inertial equation, i.e.,

$$\big|\mathbf{r}_2^\top \mathbf{S}(\gamma_2, \gamma_3)\mathbf{r}_3\big| < \varepsilon_\perp \big\|\mathbf{S}(\gamma_2, \gamma_3)\big\|_1, \tag{95}$$

where $\gamma_2 = \max(\gamma^{l_1}, \gamma^{l_2})$, $\gamma_3 = \min[\gamma^{l_1}, \gamma^{l_2}]$ and $\mathbf{r}_2$, $\mathbf{r}_3$ are the corresponding vectors from the list L_T. Each pair of inertial pairs that pass the last two tests is then appended, together with $\{\gamma_1, \mathbf{r}_1\}$, to a list $L_I = \{\gamma_1^i, \mathbf{r}_1^i; \gamma_2^i, \mathbf{r}_2^i; \gamma_3^i, \mathbf{r}_3^i\}$ of initial (approximate) solutions for Newton refinement.

In the third stage of the search, one of the two versions of Newton's method (presented in the sections 5.2 and 5.3) for solving the inertial equations is run on each of the initial solutions in the list L_I. The performance obtained with both of these methods is comparable, but the method based on the Cayley parameters is a bit faster and easier to implement, and is the method used in the examples presented in the next section.

In our implementation Newton's method we have imposed a damping factor on the step size during the first few iterations to reduce the frequency of diverging runs. Each time this iteration converges to a solution, we convert the solution into the corresponding critical matrix $\mathbf{Y}$ via equation (25), and then evaluate $f(\mathbf{Y})$ via equation (15). If this value is not already on the final list L_C, it is inserted there. The critical matrix with the smallest value of the *STRAIN* is taken as the global minimum.

It may happen that the initial γ-discretizations and the choice of $\varepsilon_\perp$ did not result in any initial solution, or resulted in too many. Furthermore, it may happen that Newton's method did not converge for any initial point, or converged too often to the same solution. These cases are controlled by starting with a relatively large $\varepsilon_\perp$ and a relatively course discretization density for the γ-mesh. Then, either $\varepsilon_\perp$ is reduced, or the γ-mesh refined, or both, to ensure finding the global minimum without an excessive amount of redundant computation.

7. Numerical experiments

The code implementing the global search procedure described in this paper was developed using the interactive MATLAB® numerical linear algebra system, and then converted into an independent 'C' program for a significantly faster execution. We have evaluated the resulting program on two kinds of test problems, one using randomly generated coordinates, and the other generated from the coordinates of structures in the Protein Data Bank (PDB). The procedures used to generate these test problems, and the results obtained with them, are the subject of this section.

7.1. Random test problems

The procedure used to generate random test problems is as follows: First, the coordinates of $\mathbf{X} \in \mathbb{R}^{M\times 3}$ and $\mathbf{Y} \in \mathbb{R}^{N\times 3}$ are chosen with a uniform distribution from within two balls centered at $\mathbf{c}_X, \mathbf{c}_Y \in \mathbb{R}^3$ with radii $r_X, r_Y > 0$. Second, both sets of points are translated to the centroid of $\mathbf{X}$ by subtracting the vector $\bar{\mathbf{x}} = -M^{-1}\sum_i \mathbf{x}_i$. Third, the inertial tensor of the $\mathbf{X}$ coordinates is diagonalized as $\mathbf{X}^\top\mathbf{X} = \mathbf{L}\mathbf{Z}\mathbf{L}^\top$ ($\mathbf{Z} = \mathrm{diag}(\zeta_1, \zeta_2, \zeta_3)$ and $\zeta_1 \geqslant \zeta_2 \geqslant \zeta_3$), and both the $\mathbf{X}$ and $\mathbf{Y}$ coordinates are rotated by $\mathbf{L}$ so as to obtain principal axis coordinates for $\mathbf{X}$. Finally, the coordinates of both

$\mathbf{X}$ and $\mathbf{Y}$ are scaled by $\zeta_1^{-1/2}$, so as to equalize the norm of $\mathbf{X}^\top\mathbf{X}$ across different problems.

The matrices of squared distances, i.e.,

$$\begin{aligned}\mathbf{D}^A &= \big[D^A_{ij}\big]^{M,M}_{i,j=1} = \big[\|\mathbf{x}_i - \mathbf{x}_j\|^2\big]^{M,M}_{i,j=1},\\ \mathbf{D}^B &= \big[D^B_{ij}\big]^{M,N}_{i,j=1} = \big[\|\mathbf{x}_i - \mathbf{y}_j\|^2\big]^{M,N}_{i,j=1},\\ \mathbf{D}^C &= \big[D^C_{ij}\big]^{N,N}_{i,j=1} = \big[\|\mathbf{y}_i - \mathbf{y}_j\|^2\big]^{N,N}_{i,j=1},\end{aligned} \tag{96}$$

are then calculated from these coordinates, and their elements perturbed by uniformly distributed random numbers $1 + \vartheta_{ij}$ as follows:

$$\begin{aligned}\widetilde{\mathbf{D}}^B &= \big[\widetilde{D}^B_{ij}\big]^{M,N}_{i,j=1} \equiv \big[(1+\vartheta_{ij})^2 D^B_{ij}\big]^{M,N}_{i,j=1} \quad \big(\vartheta_{ij} \in (-\varepsilon_B, \varepsilon_B)\big),\\ \widetilde{\mathbf{D}}^C &= \big[\widetilde{D}^C_{ij}\big]^{N,N}_{i,j=1} \equiv \big[(1+\vartheta_{ij})^2 D^C_{ij}\big]^{N,N}_{i,j=1} \quad \big(\vartheta_{ij} \in (-\varepsilon_C, \varepsilon_C)\big).\end{aligned} \tag{97}$$

We stress that these perturbed distance matrices can no longer be embedded in a three-dimensional Euclidean space.

In order to have the blocks of the corresponding Gram matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ refer to a common origin and at the same time ensure that $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$ is fit exactly, it is necessary place the origin in the $\mathbf{X}$ set, preferably at its centroid as above. This can be done directly from the distances by assigning unit mass to all the $\mathbf{X}$ coordinates, and zero to the $\mathbf{Y}$ coordinates. Equation (4) then yields

$$\begin{aligned}D^A_{0i} &= \frac{1}{M}\sum_{j=1}^{M} D^A_{ij} - \frac{1}{M^2}\sum_{1=j<k}^{M,M} D^A_{jk} \quad (i = 1, \ldots, M),\\ \widetilde{D}^C_{0i} &= \frac{1}{M}\sum_{j=1}^{M} \widetilde{D}^B_{ij} - \frac{1}{M^2}\sum_{1=j<k}^{M,M} D^A_{jk} \quad (i = 1, \ldots, N).\end{aligned} \tag{98}$$

Finally, we set $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$ and

$$\begin{aligned}\mathbf{B} &= \big[\big(D^A_{0i} + \widetilde{D}^C_{0j} - \widetilde{D}^B_{ij}\big)/2\big]^{M,N}_{i,j=1},\\ \mathbf{C} &= \big[\big(\widetilde{D}^C_{0i} + \widetilde{D}^C_{0j} - \widetilde{D}^C_{ij}\big)/2\big]^{N,N}_{i,j=1}.\end{aligned} \tag{99}$$

Thus the free parameters that define our test problems include the dimensions M and N of $\mathbf{X}$ and $\mathbf{Y}$, the ratio of distance between the centers of the spheres to the sum of their radii, $s_{XY} = \|\mathbf{c}_X - \mathbf{c}_Y\|/(r_X + r_Y)$, and the perturbation magnitudes ε_B and ε_C. In the problems reported below, we set $\{M = 10, N = 50\}$, $\{M = N = 30\}$ and $\{M = 50, N = 10\}$ with $r_X = M^{1/3}$ and $r_Y = N^{1/3}$, so as to keep the density approximately constant in each case. In addition, we set $s_{XY} = 0$, $s_{XY} = 1/2$ and $s_{XY} = 1$ for each of these three cases, as well as $\varepsilon_B = \varepsilon_C = \varepsilon$ for three perturbation levels, $\varepsilon = 0.005, 0.05, 0.5$. We generate and solve ten random problems in each class,

Table 1
Statistics for random test problems (see text).

ε (%)	M, N	s_{XY}	(a)	(b)	(c)	(d)	(e)	(e)	(g)
0.5	50, 10	0.0	8, 29	58	5836	4959	1.0	0.0000	672
0.5	50, 10	0.5	22, 45	219	2234	1270	1.0	0.0002	315
0.5	50, 10	1.0	22, 68	199	4136	2010	1.0	0.0007	554
0.5	30, 30	0.0	25, 43	218	2563	623	1.0	0.0003	485
0.5	30, 30	0.5	21, 72	161	4026	931	1.1	0.0012	615
0.5	30, 30	1.0	21, 71	140	5001	641	1.1	0.0077	515
0.5	10, 50	0.0	22, 91	70	6532	743	3.2	0.0248	838
0.5	10, 50	0.5	21, 90	43	6755	725	4.2	0.0514	813
0.5	10, 50	1.0	21, 84	35	6817	874	2.4	0.1443	1018
5.0	50, 10	0.0	12, 31	65	5094	4144	1.0	0.0022	533
5.0	50, 10	0.5	25, 46	290	766	439	1.0	0.0207	127
5.0	50, 10	1.0	24, 69	253	2431	1180	1.1	0.0726	303
5.0	30, 30	0.0	33, 43	320	1984	514	1.0	0.0273	410
5.0	30, 30	0.5	24, 83	183	5997	650	1.2	0.1340	803
5.0	30, 30	1.0	24, 86	152	5587	881	1.2	0.8100	599
5.0	10, 50	0.0	24, 96	146	5828	430	3.2	2.1188	698
5.0	10, 50	0.5	24, 89	76	6123	880	2.9	7.7206	699
5.0	10, 50	1.0	24, 95	47	6246	731	2.8	10.038	821
50.0	50, 10	0.0	9, 31	67	6222	4630	1.0	0.2135	713
50.0	50, 10	0.5	22, 44	221	2440	1460	1.0	1.8616	325
50.0	50, 10	1.0	22, 66	219	2989	1333	1.0	7.0796	394
50.0	30, 30	0.0	35, 49	285	3235	728	1.0	5.2963	655
50.0	30, 30	0.5	21, 70	192	5177	537	1.1	18.166	696
50.0	30, 30	1.0	21, 69	136	5771	1542	1.2	82.830	780
50.0	10, 50	0.0	21, 89	78	6395	411	3.7	203.86	759
50.0	10, 50	0.5	21, 98	38	6892	1003	3.0	471.46	842
50.0	10, 50	1.0	21, 103	45	6850	782	3.5	1292.1	901

so that the total number of random problems is $10\times3\times3\times3 = 270$. All problems were initialized with an arc length increment of $\Delta L = \pi/6$ (for the sequence of intersection curves) and an orthogonality threshold of $\varepsilon_\perp = \sqrt{1/2}\ \mathrm{asin}(\pi/6)$.

For each of the 27 classes of problems, we report the average over the 10 problems solved of: (a) the number of discretized points generated for γ_1 and for γ_2; (b) the total number of unit vectors $\mathbf{r}_1$ generated on the intersection curves for entire γ_1-mesh; (c) the number of initial points to which Newton's method was applied; (d) the number of points for which Newton's method converged; (e) the number of distinct solutions to the inertial equations located; (f) the minimum function value found; (g) the CPU time required for the entire procedure (in seconds). These statistics are summarized in table 1.

It may be seen that the number of critical matrices found (e) was usually just one, the exception being those problems in which **Y** contained more points (50) than **X** (10), where it averaged about four. The relatively small number of critical matrices satisfying bounds is of course what makes the global search feasible. We observed that

the upper bound γ^+ and the lower bound γ^- were quite tight, in that their values at the global minimum were usually within a few percent of these bounds. The upper bound γ_2^+, however, was sometimes several times too large, particularly in the $s_{XY} = 1$ cases. We also computed the standard deviations for all the averages given in table 1 (not shown). The standard deviations in the number of approximate solutions found (c) and the number of times Newton's method converged (d) were always comparable to the averages for these quantities, and since these calculations consumed most of the CPU time, the average times reported in (g) fluctuated widely even within a single class of problem.

Numerical experience so far indicates that, with a moderately small discretization step, the search procedure always succeeds in finding the global minimum of the *STRAIN*. The difficulties encountered in an earlier version of the code, which occurred when the global minimum lay in the immediate vicinity of a pole, tangency or bifurcation point, were eliminated by: (a) improvements in the discretization criterion; (b) increasing the mesh density in the vicinity of such points; (c) decreasing the orthogonality threshold $\varepsilon_\perp$ and the discretization criteria for the γ_1 and γ_2 meshes whenever a run failed to find any critical points.

7.2. Chemical test problems

For our chemical test problems, we used two biologically representative systems. The first was the crystal structure of a complex of the enzyme dihydrofolate reductase (DHFR; entry 3DRC in the Brookhaven Protein Data Bank) with its inhibitor methotrexate (MTX) [5]. Only the nonhydrogen atoms of one of the two molecules in the asymmetric unit were used, which gave 33 MTX atoms for the unknown **Y** coordinates, 1262 DHFR atoms for the fixed **X** coordinates. The second was the crystal structure of a small protein, bovine pancreatic trypsin inhibitor (BPTI; entry 4PTI in the Protein Data Bank), once again without hydrogen atoms [10]. In this case the coordinates of the 284 backbone nitrogen, alpha-carbon, beta-carbon, carbonyl-carbon and carbonyl-oxygen atoms were used for the **X** set, while the remaining 170 sidechain atoms were used for the **Y** set.

In both cases the distance matrices were calculated from the coordinates in the crystal structures, and those distances involving the chosen **Y** coordinates were perturbed. The Gram matrices **B** and **C** were generated from these perturbed distances exactly as described for the random test problems above, but a somewhat different procedure was used to perturb the distances. First, if the distance d was less than three Ångstroms, it was used unchanged. Second, if the distance d was greater than three Ångstroms, it was replaced by a uniform random number between $\max(3.0, d/\sqrt{2})$ and $d\sqrt{2}$. In the case of BPTI, this procedure generates random distance matrices like those that would be used as input to the *EMBED* algorithm in a homology modeling problem wherein only the backbone conformation of the protein was known. In the case of DHFR/MTX, the procedure generates random distance matrices like those that would be available if the structure of the protein was already known, and a number of

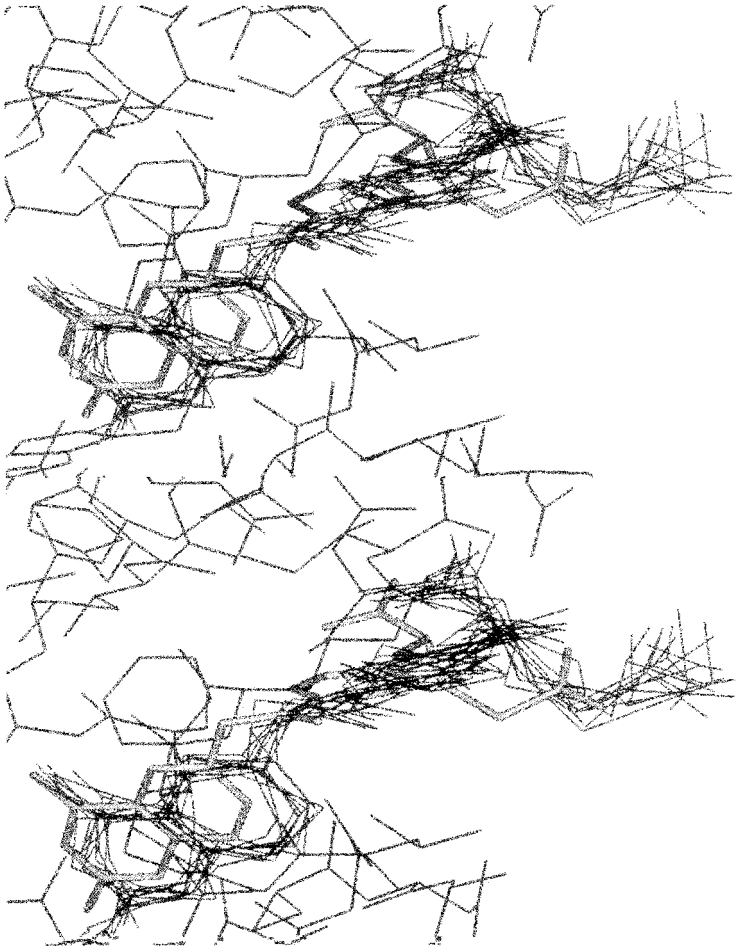

Figure 5. Stereoview of the results of the DHFR/MTX test problems. The 10 MTX structures generated are drawn with medium lines, while the DHFR protein is drawn with thin lines, and the MTX crystal structure is a shaded stick drawing.

ligand-protein and ligand-ligand interatomic contacts had been independently identified from NMR data.

The total number of random perturbations solved was 10 for both the DHFR/MTX and the BPTI backbone/sidechains problems. The DHFR/MTX trials ran smoothly, and although the number of approximate solutions found exceeded 6000 in all 10 trials, Newton's method converged better than 95% of the time to the same critical matrix every time, which is therefore almost certainly the unique global minimum. A single unique global minimum was also found in every case for BPTI, although Newton's method failed to converge much more often. This is due to the fact that γ_2 was always close to the bifurcation point γ_b, and the angle between the curves at γ_b was always very small (cf. section 4.1). The time required for the BPTI problems ranged from 3

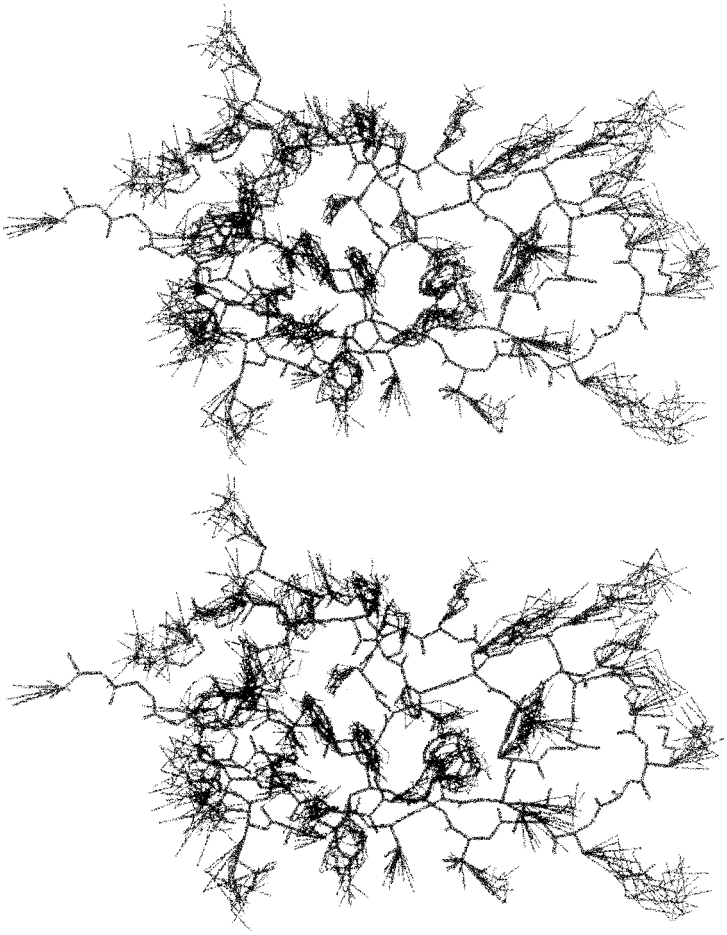

Figure 6. Stereoview of the results of the BPTI test problems treating the backbone atoms as fixed and the sidechain atoms as variable. The backbones of the computed structures have been superimposed on that of the crystal structure, which is drawn with a heavy line for comparison.

to 15 minutes.

Figures 5 and 6 show the results of the DHFR/MTX and BPTI test problems. Although a certain scatter is apparent in these coordinates, they are all very close to the crystal structure coordinates from which the unperturbed distances were obtained. This ability to correct for large (ca. 40%) but random perturbations in the distances is what makes the minimization criterion (9) and the algorithm a powerful tool for determination of structures. The figures also show that the perturbation procedure introduced some bias away from the crystal structure, which is expected when a simple uniform distribution is used for the perturbed distances [15]. How the procedure performs with the estimated distances that are available in actual applications will be the subject of a further study.

8. Closing comments

In this paper we have given a precise mathematical statement of the problem of embedding with a rigid substructure, derived necessary conditions for the global minimum, developed a search strategy for finding it, and evaluated the resulting algorithm on a realistic set of test problems. The rigid embedding problem, however, is only one step in a longer sequence of calculations by which one generates atomic coordinates that satisfy distance constraints. This means that in actual applications it will be necessary to have methods of incorporating fixed atom constraints into the other steps of the overall calculation. These include bound smoothing, the generation of approximate distances from which to build the metric matrix for embedding, and the optimization of the resulting coordinates.

It turns out that these steps are all relatively straightforward. Fixed atoms are built into the bound smoothing and distance generation steps simply by setting the distances among the fixed atoms to their known values. Similarly, the fixed atom constraints are incorporated into the optimization simply by eliminating their coordinates from the list of variables to be used. A suite of programs derived from the *DG-II* distance geometry program package [15] which incorporate these modifications will be described elsewhere, along with their application to chemically and biologically important problems [18].

Acknowledgements

We thank Sven Hyberts for help with the chemical test problems. This work was supported by NSF grants BIR-9511892 and MCB-9527181, by NIH grants GM-38221 and GM-47467, and by a grant from Molecular Simulations, Inc.

References

[1] S.L. Altmann, *Rotations, Quaternions and Double Groups* (Clarendon Press, Oxford, 1986).
[2] A.L. Andrew and K.-W. Eric Chu, SIAM J. Matrix Anal. Appl. 14 (1993) 903–926.
[3] R. Bellman, *Introduction to Matrix Analysis* (reprinted by SIAM, Philadelphia, 1995).
[4] J.M. Blaney and J.S. Dixon, in: *Reviews in Computational Chemistry*, Vol. V, eds. K.B. Lipkowitz and D.B. Boyd (VCH Publishers, New York, 1994) pp. 299–335.
[5] J.T. Bolin, D.J. Filman, D.A. Matthews, R.C. Hamlin and J. Kraut, J. Biol. Chem. 257 (1982) 13650–13662.
[6] C. Cohen-Tannoudji, B. Diu and F. Laloë, *Quantum Mechanics* (Wiley, New York, 1977) p. 1192.
[7] G.M. Crippen, J. Math. Chem. 6 (1991) 307–324.
[8] G.M. Crippen and T.F. Havel, Acta Cryst. A34 (1978) 282–284.
[9] G.M. Crippen and T.F. Havel, *Distance Geometry and Molecular Conformation* (Research Studies Press, Taunton, 1988).
[10] J. Deisenhofer and W. Steigemann, Acta Cryst. B31 (1975) 238–240.
[11] C. Eckart and G. Young, Psychometrika 1 (1936) 211–218.
[12] F.R. Gantmacher, *The Theory of Matrices* (Chelsea, Bronx, 1959).
[13] J.C. Gower, Lin. Alg. Appl. 67 (1985) 81–97.

[14] A. Graham, *Kronecker Products and Matrix Calculus with Applications* (Wiley, New York, 1981).
[15] T.F. Havel, Prog. Biophys. Mol. Biol. 56 (1991) 43–78.
[16] T.F. Havel, Mol. Simul. 10 (1993) 175–210.
[17] T.F. Havel, in: *Encyclopedia of Nuclear Magnetic Resonance*, eds. D.M. Grant and R.K. Harris (Wiley, New York, 1996) pp. 1701–1710.
[18] T.F. Havel, S. Hyberts and I. Najfeld, German Conf. Bioinformatics, 1996 (to be published in Lect. Notes in Comput. Sci.).
[19] R.A. Horn and C.R. Johnson, *Topics in Matrix Analysis* (Cambridge Univ. Press, Cambridge, 1991).
[20] A.S. Householder, *The Numerical Treatment of a Single Nonlinear Equation* (McGraw-Hill, New York, 1970).
[21] J. Meulman, *A Distance Approach to Nonlinear Multivariate Analysis* (DSWO Press, Leiden, 1986).
[22] L.B. Rall, *Computational Solution of Nonlinear Operator Equations* (Wiley, New York, 1969).