



Sensitivity analysis of the strain criterion for multidimensional scaling

R.M. Lewis, M.W. Trosset*

Department of Mathematics, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23185-8795, USA

Available online 18 August 2004

Abstract

Multidimensional scaling (MDS) is a collection of data analytic techniques for constructing configurations of points from dissimilarity information about interpoint distances. Classical MDS assumes a fixed matrix of dissimilarities. However, in some applications, e.g., the problem of inferring 3-dimensional molecular structure from bounds on interatomic distances, the dissimilarities are free to vary, resulting in optimization problems with a spectral objective function. A perturbation analysis is used to compute first- and second-order directional derivatives of this function. The gradient and Hessian are then inferred as representers of the derivatives. This coordinate-free approach reveals the matrix structure of the objective and facilitates writing customized optimization software. Also analyzed is the spectrum of the Hessian of the objective.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Classical multidimensional scaling; Principal coordinate analysis; Distance matrices; Distance geometry; Spectral decomposition; Perturbation analysis

1. Introduction

Multidimensional scaling (MDS) is a collection of data analytic techniques for constructing configurations of points from dissimilarity information about interpoint distances. Developed primarily by psychometricians and statisticians, MDS is widely used in a variety of disciplines for visualization and dimension reduction. The extensive literature on MDS

* Corresponding author. Tel.: +1-757-221-2040; fax: +1-757-221-7400.

E-mail addresses: buckaroo@math.wm.edu (R.M. Lewis), trosset@math.wm.edu (M.W. Trosset)

URLs: <http://www.math.wm.edu/~buckaroo>, <http://www.math.wm.edu/~trosset>.

includes numerous books, e.g., Borg and Groenen (1997), Cox and Cox (1994), and Everitt and Rabe-Hesketh (1997); and book chapters, e.g., Everitt and Dunn (1991, Chapter 5), Krzanowski and Marriott (1994, Chapter 5), Mardia et al. (1979, Chapter 14), and Seber (1984, Section 5.5). Many specific formulations of MDS are possible; a useful organizing principle, adopted by de Leeuw and Heiser (1982) and by Trosset (1997b), is to formulate MDS as a collection of optimization problems. The method of Torgerson (1952) and Gower (1966), variously called classical MDS and principal coordinate analysis, can be formulated as an optimization problem with an objective function whose minimum value is sometimes called the strain criterion.

Formally, a matrix is a *dissimilarity matrix* if and only if it is symmetric and hollow (its diagonal entries vanish) with nonnegative entries. Given an $n \times n$ matrix $\Delta = (\delta_{ij})$ of squared dissimilarities and a target dimension p , a classical problem in distance geometry is to determine if Δ is a matrix of p -dimensional squared Euclidean distances, i.e., if there exist $x_1, \dots, x_n \in \mathbb{R}^p$ such that $\|x_i - x_j\|^2 = \delta_{ij}$. It is well-known that the answer is affirmative if and only if the symmetric $n \times n$ matrix

$$A = \tau(\Delta) = -\frac{1}{2} P^T \Delta P$$

is positive semidefinite with $\text{rank}(A) \leq p$, where P is the $n \times n$ projection matrix $I - ee^T/n$, I is the $n \times n$ identity matrix, and $e = (1, \dots, 1)^T \in \mathbb{R}^n$. This embedding theorem motivated classical MDS, which can be stated as the problem of finding the symmetric positive semidefinite $n \times n$ matrix of rank $\leq p$ that is nearest $\tau(\Delta)$ in squared Frobenius distance. The minimum value of the objective function, $\|B - \tau(\Delta)\|_F^2$, is the strain criterion. Detailed studies of the linear operator τ were made by Critchley (1988) and by Gower and Groenen (1991).

To evaluate the strain criterion for a fixed Δ , one computes the spectral decomposition $\tau(\Delta) = Q\Lambda Q^T$ and sets $\bar{\Lambda} = \text{diag}(\bar{\lambda})$, where

$$\bar{\lambda}_i = \begin{cases} \max(\lambda_i, 0) & i = 1, \dots, p \\ 0 & i = p + 1, \dots, n \end{cases}$$

and $\text{diag}(\bar{\lambda})$ is the diagonal matrix whose diagonal entries are $(\bar{\lambda}_1, \dots, \bar{\lambda}_n)$. Then $Q\bar{\Lambda}Q^T$ is a global minimizer of $\|B - \tau(\Delta)\|_F^2$ and the global minimum is

$$\begin{aligned} F_p \circ \tau(\Delta) &= \|Q\bar{\Lambda}Q^T - Q\Lambda Q^T\|_F^2 = \sum_{i=1}^n (\bar{\lambda}_i - \lambda_i)^2 \\ &= \sum_{i=1}^p [\max(\lambda_i, 0) - \lambda_i]^2 + \sum_{i=p+1}^n \lambda_i^2 = \sum_{i=1}^n \lambda_i^2 - \sum_{i=1}^p [\max(\lambda_i, 0)]^2 \\ &= \|\tau(\Delta)\|_F^2 - \sum_{i=1}^p [\max(\lambda_i, 0)]^2. \end{aligned} \quad (1)$$

Notice that only the p largest eigenvalues are required to evaluate the strain criterion.

Classical MDS assumes that Δ is fixed. Recently, extensions of classical MDS have been developed for nonmetric MDS (Trosset, 1998b) and for various problems with bound constraints (Trosset, 1998a, 2000). The latter include the important problem of inferring

3-dimensional molecular structure from partial information about interatomic distances, which is especially challenging because the number of atoms (n) is typically large and precise solutions are required. Trosset (2002) analyzed the computational theory of these extensions and reported computational results for several molecular conformation problems. To efficiently minimize $F_3 \circ \tau(\Delta)$ (for simplicity, we henceforth restrict attention to $p = 3$) as Δ varies, we require first- and second-order derivative information about the objective function $F_3 \circ \tau$.

First derivatives of $F_3 \circ \tau$ can be derived by various methods. One useful technique involves perturbation analysis, as in Wilkinson (1965). This technique was exploited by Sibson (1979) in his study of the robustness of classical MDS. Another technique involves the theory of reducible nonlinear programming, as in Parks (1985); yet another technique involves the theory of spectral functions, as in Lewis (1996). These techniques were exploited by Trosset (1997a, 2002) to derive the first-order partial derivatives of $F_3 \circ \tau$ with respect to the components of Δ .

Previous research has not considered the second derivatives of $F_3 \circ \tau$. Noting that there are $O(n^4)$ second-order partial derivatives of $F_3 \circ \tau$, Trosset (2002) opted to use a limited memory quasi-Newton method for optimization. The primary purpose of the present work is to derive second derivatives of $F_3 \circ \tau$. We do so via perturbation analysis; thus, our preliminary derivation of first derivatives overlaps Sibson (1979). The essence of our approach lies in writing the second-order Taylor polynomial approximation of $F_3 \circ \tau$ in terms of duality pairings of symmetric matrices via the Frobenius inner product. We then compute directional derivatives of $F_3 \circ \tau$ and infer the gradient and Hessian as representers of the derivatives. This allows us to manipulate matrices directly, rather than working with matrix components. The result is a more lucid, coordinate-free approach that better reveals the matrix structure of the objective function and better lends itself to writing customized software for minimizing $F_3 \circ \tau$. Furthermore, because we compute Hessian-matrix pairings, we can avail ourselves of optimization algorithms based on iterative linear algebra, and we circumvent the memory requirements that dissuaded Trosset (2002) from using second-order methods. We conclude by analyzing the spectrum of the Hessian of $F_3 \circ \tau$.

2. Some properties of the Frobenius inner product

For later reference, we recall some properties of the Frobenius inner product. We are interested here only in the case of real $n \times n$ matrices, for which the Frobenius inner product is

$$\langle A, B \rangle_F = \text{trace}(A^T B) = \sum_{i=1}^n \sum_{j=1}^n A_{ij} B_{ij}.$$

For any $n \times n$ matrix F , we have

$$\langle A, F^T B F \rangle_F = \text{trace}(A^T F^T B F) = \text{trace}(F A^T F^T B) = \langle F A F^T, B \rangle_F.$$

Also note that for any vectors $u = (u_1, \dots, u_n)^T$ and $v = (v_1, \dots, v_n)^T$,

$$u^T A v = \sum_{i=1}^n \sum_{j=1}^n u_i A_{ij} v_j = \left\langle A, u v^T \right\rangle_F = \left\langle A, \frac{1}{2}(u v^T + v u^T) \right\rangle_F. \quad (2)$$

3. Differentiability of the strain criterion

Before computing the first and second derivatives of $F_3 \circ \tau$, we consider where it is first- and second-order differentiable. To do so, we rely on results from the theory of spectral functions. Given a symmetric matrix T , let $\lambda(T) \in \mathbb{R}^n$ denote the eigenvalues of T . A real-valued function of symmetric matrices, G , is a spectral function if and only if we can write $G(T) = g(\lambda(T))$, where $g : \mathbb{R}^n \rightarrow \mathbb{R}$ is symmetric, i.e., $g(\mu) = g(P\mu)$ for every permutation matrix P .

Lemma 3.1. *Let $g : \mathbb{R}^n \rightarrow \mathbb{R}$ be symmetric and define a spectral function G by $G(T) = g(\lambda(T))$.*

- (1) (Lewis, 1996, Theorem 1.1) *G is differentiable at T if and only if g is differentiable at $\lambda(T)$. If G is differentiable at T , then*

$$\nabla G(T) = U^T (\text{diag}[\nabla g(\lambda(T))]) U, \quad (3)$$

where U is any unitary matrix for which $T = U^T (\text{diag}(\lambda(T))) U$.

- (2) (Lewis and Sendov, 2001, Theorem 3.3) *G is twice continuously differentiable at T if and only if g is twice continuously differentiable at $\lambda(T)$.*

Because τ is linear, $F_3 \circ \tau$ is differentiable at Δ if and only if F_3 is differentiable at $T = \tau(\Delta)$. We apply Lemma 3.1 to F_3 . Given $\mu = (\mu_1, \dots, \mu_n)^T$, let

$$\mu_{(1)} \geq \mu_{(2)} \geq \mu_{(3)} \geq \dots \geq \mu_{(n)}$$

denote the order statistics of μ and define a symmetric function g by

$$g(\mu) = \sum_{i=1}^n \mu_i^2 - \sum_{i=1}^3 [\max(\mu_{(i)}, 0)]^2.$$

Then $F_3(T) = g(\lambda(T))$, so F_3 is a spectral function and we have the following:

Theorem 3.1. *Let $F_3 \circ \tau$ be the function defined by (1). Given Δ , let $\lambda_1 \geq \dots \geq \lambda_n$ denote the eigenvalues of $\tau(\Delta)$.*

- (1) (Trosset, 2002, Theorem 4) *$F_3 \circ \tau$ is differentiable at Δ , unless $\lambda_3 = \lambda_4 \geq 0$.*
 (2) *$F_3 \circ \tau$ is twice continuously differentiable at Δ , unless $\lambda_3 = \lambda_4 \geq 0$ or at least one of $\lambda_1, \lambda_2, \lambda_3$ vanishes.*

If Δ is a matrix of squared Euclidean distances, then all of the eigenvalues of $\tau(\Delta)$ are nonnegative and the conditions in Theorem 3.1 have geometric interpretations. The

condition $\lambda_3 = \lambda_4$ obtains if and only if the configuration of points that generate Δ has equal variation in the third and fourth dimensions. If Δ is three-dimensional but not two-dimensional, then this is impossible. Furthermore, if Δ is three-dimensional but not two-dimensional, then $\lambda_1 \geq \lambda_2 \geq \lambda_3 > 0$.

4. The gradient and Hessian of the strain criterion

Now we derive the gradient and Hessian of the function $F = F_3 \circ \tau$. We first study $F = F(D)$ as a function of general symmetric matrices, later restricting attention to hollow symmetric matrices. In the course of these derivations, we use δ to denote a perturbation (variation). The derivation that we present here assumes that the three largest eigenvalues of $\tau(D)$ are each simple. However, we argue later that the formulae hold even if some of these eigenvalues coalesce.

We derive the action of the derivatives as linear maps of $n \times n$ symmetric matrices, and from these actions infer the identity of these maps, in the sense of identifying Riesz representers for these maps with respect to the Frobenius inner product. Because we use the Frobenius inner product, the second-order Taylor's series expansion of F is

$$F(D + \delta D) = F(D) + \langle \nabla F(D), \delta D \rangle_F + \frac{1}{2} \langle \nabla^2 F(D) \delta D, \delta D \rangle_F + \text{h.o.t.},$$

where h.o.t. denotes the higher order (remainder) terms in the expansion. Identifying the gradient and the Hessian as Riesz representers contrasts with Trosset's (2002) approach of computing partial derivatives with respect to individual entries in the matrix of squared dissimilarities. Our representation of the derivatives in terms of matrices, rather than entries of matrices, greatly simplifies both the task of analysis and the task of implementing an algorithm to minimize $F_3 \circ \tau$. Mathematically, our approach is appealing because it uses the natural structure of the problem.

We compute the first and second derivatives of $F = F_3 \circ \tau$ in several steps. First, we compute the derivatives with respect to a symmetric matrix A of a single eigenvalue–eigendirection pair $(\lambda(A), v(A))$ of A . Then we compute the derivatives of λ^2 , then those of $\lambda(\tau(D))$, and finally the derivatives of $F = F_3 \circ \tau$. Our derivation is *not* based on the formulas derived by Lewis (1996) and Lewis and Sendov (2001) for the first and second derivatives of spectral functions. We can apply (3) to F_3 and obtain $\nabla F_3 \circ \tau$ in short order, as did Trosset (2002). However, applying the expression in Lewis and Sendov (2001) for the Hessian leads to a quantity whose identity is obscure and difficult to clarify. We believe that our derivation leads to a characterization of the Hessian that is easier to understand.

4.1. Derivatives of simple eigenvalue–eigendirection pairs

Let A be a real symmetric $n \times n$ matrix. An eigenvalue–eigendirection pair (v, λ) for A is the solution of the system

$$\begin{aligned} (A - \lambda I)v &= 0, \\ \frac{1}{2} v^T v &= \frac{1}{2}. \end{aligned} \tag{4}$$

The Jacobian of this system with respect to v and λ is

$$\begin{pmatrix} A - \lambda I & v \\ v^T & 0 \end{pmatrix}.$$

If λ is a simple eigenvalue, then the Jacobian is invertible and, by the implicit function theorem, (v, λ) is locally a smooth function of the entries of A .

The sensitivity analysis for the first derivatives of λ is classical, e.g., [Wilkinson \(1965\)](#). Our use of it replicates [Sibson's \(1979\)](#) study of the robustness of classical MDS. However, as we need first derivatives to calculate second derivatives, we include the details of the first-order analysis.

We have

$$\begin{aligned} \lambda(A + \delta A) &= \lambda(A) + \langle \nabla \lambda(A), \delta A \rangle_F + \frac{1}{2} \langle \delta A, \nabla^2 \lambda(A) \cdot \delta A \rangle_F + \text{h.o.t.}, \\ v(A + \delta A) &= v(A) + Dv(A) \cdot \delta A + \frac{1}{2} (D^2 v(A) \cdot \delta A) \delta A + \text{h.o.t.}, \end{aligned}$$

where h.o.t. again denotes the higher order (remainder) terms in the Taylor's expansion. For brevity, we introduce the notation

$$\begin{aligned} \delta \lambda &= \langle \nabla \lambda(A), \delta A \rangle_F, & \delta^2 \lambda &= \frac{1}{2} \langle \delta A, \nabla^2 \lambda(A) \cdot \delta A \rangle_F, \\ \delta v &= Dv(A) \cdot \delta A, & \delta^2 v &= \frac{1}{2} (D^2 v(A) \cdot \delta A) \delta A, \end{aligned}$$

to denote the first- and second-order effects of the perturbation δA .

Note that $\lambda(A + \delta A)$ and $v(A + \delta A)$ also satisfy (4), whence

$$\begin{aligned} ((A + \delta A) - (\lambda + \delta \lambda + \delta^2 \lambda + \text{h.o.t.})I)(v + \delta v + \delta^2 v + \text{h.o.t.}) &= 0, \\ \frac{1}{2} (v + \delta v + \delta^2 v + \text{h.o.t.})^T (v + \delta v + \delta^2 v + \text{h.o.t.}) &= \frac{1}{2}. \end{aligned}$$

Expanding and grouping terms of like order, and applying (4), we obtain

$$(A - \lambda I) \delta v + (\delta A - \delta \lambda I) v = 0, \tag{5}$$

$$v^T \delta v = 0 \tag{6}$$

and

$$(A - \lambda I) \delta^2 v + (\delta A - \delta \lambda I) \delta v - \delta^2 \lambda v = 0, \tag{7}$$

$$v^T \delta^2 v + \frac{1}{2} \delta v^T \delta v = 0. \tag{8}$$

Multiplying (5) by v^T and using the fact that v is a eigenvector of length 1, we obtain

$$v^T (A - \lambda I) \delta v + v^T (\delta A - \delta \lambda I) v = v^T \delta A v - \delta \lambda = 0,$$

or

$$\delta \lambda = v^T \delta A v.$$

From (2) we see that

$$\delta \lambda = \left\langle v v^T, \delta A \right\rangle_F,$$

from which we conclude that $\nabla \lambda$ the Riesz representer (in the Frobenius inner product) for the first derivative of λ with respect to A , is

$$\nabla \lambda = vv^T. \tag{9}$$

As noted previously, this also follows immediately from (3).

Meanwhile, from (5) we see that

$$\delta v = -(A - \lambda I)^+(\delta A - \delta \lambda I)v. \tag{10}$$

Because v is the eigenvector associated with λ , we have

$$(A - \lambda I)^+v = 0, \tag{11}$$

so (10) simplifies to

$$\delta v = -(A - \lambda I)^+\delta Av. \tag{12}$$

To compute the second variations $\delta^2 \lambda$ and $\delta^2 v$, we first multiply (7) by v^T :

$$\begin{aligned} v^T(A - \lambda I)\delta^2 v + v^T(\delta A - \delta \lambda I)\delta v - v^T\delta^2 \lambda v \\ = v^T(\delta A - \delta \lambda I)\delta v - \delta^2 \lambda = 0, \end{aligned}$$

whence

$$\delta^2 \lambda = v^T(\delta A - \delta \lambda I)\delta v.$$

Appealing to (12) we obtain

$$\delta^2 \lambda = -v^T(\delta A - \delta \lambda I)(A - \lambda I)^+\delta Av = -v^T\delta A(A - \lambda I)^+\delta Av.$$

Let

$$u = (A - \lambda I)^+\delta Av. \tag{13}$$

Then

$$\delta^2 \lambda = -v^T\delta Au,$$

and by (2),

$$\delta^2 \lambda = -\left\langle \delta A, \frac{1}{2}(uv^T + vu^T) \right\rangle_F.$$

From the definition of $\delta^2 \lambda$ we conclude that

$$\begin{aligned} \nabla^2 \lambda \cdot \delta A &= -(uv^T + vu^T) \\ &= -\left[(A - \lambda I)^+\delta Avv^T + vv^T\delta A(A - \lambda I)^+ \right]. \end{aligned} \tag{14}$$

4.2. The chain rule

Next we turn to the derivatives of $\lambda^2(\tau(D))$. Here, D denotes a general symmetric matrix, not necessarily a matrix of squared dissimilarities. In particular, we do not assume that D is hollow.

Instead of attempting a cumbersome application of a general chain rule, we derive the specific chain rule that we need from first principles. We compute the derivatives in two steps. First, we account for the effect of squaring the eigenvalue. Let $\phi(A) = \lambda^2(A)$. Then

$$\begin{aligned}\phi(A + \delta A) &= (\lambda + \delta\lambda + \delta^2\lambda + \text{h.o.t.})^2 \\ &= \lambda^2 + 2\lambda\delta\lambda + (\delta\lambda)^2 + 2\lambda\delta^2\lambda + \text{h.o.t.} \\ &= \lambda^2 + 2\lambda\delta\lambda + \frac{1}{2}(2(\delta\lambda)^2 + 4\lambda\delta^2\lambda) + \text{h.o.t.},\end{aligned}$$

so

$$\langle \nabla\phi, \delta A \rangle_F = 2\lambda\delta\lambda = 2\lambda\langle \nabla\lambda, \delta A \rangle_F,$$

or, from (9),

$$\nabla\phi = 2\lambda\nabla\lambda = 2\lambda uv^T, \quad (15)$$

as we would expect. We also see that

$$\begin{aligned}\langle \delta A, \nabla^2\phi \cdot \delta A \rangle_F &= 2(\delta\lambda)^2 + 4\lambda\delta^2\lambda \\ &= 2(v^T\delta Av)(v^T\delta Av) + 4\lambda\frac{1}{2}\langle \delta A, \nabla^2\lambda \cdot \delta A \rangle_F \\ &= 2v^T\delta Avv^T\delta Av + 2\lambda\langle \delta A, \nabla^2\lambda \cdot \delta A \rangle_F \\ &= 2\langle \delta A, vv^T\delta Avv^T \rangle_F + 2\lambda\langle \delta A, \nabla^2\lambda \cdot \delta A \rangle_F,\end{aligned}$$

so by (14),

$$\begin{aligned}\nabla^2\phi \cdot \delta A &= 2vv^T\delta Avv^T + 2\lambda\nabla^2\lambda \cdot \delta A \\ &= 2vv^T\delta Avv^T + 2\lambda(uv^T + vu^T).\end{aligned} \quad (16)$$

Finally, we come to $\Phi(D) = \phi(\tau(D)) = \lambda^2(\tau(D))$. We have

$$\begin{aligned}\Phi(D + \delta D) &= \phi(-\frac{1}{2}P^T(D + \delta D)P) \\ &= \phi(\tau(D) - \frac{1}{2}P^T\delta DP) \\ &= \phi(\tau(D)) + \langle \nabla\phi(\tau(D)), -\frac{1}{2}P^T\delta DP \rangle_F \\ &\quad + \frac{1}{2}\langle -\frac{1}{2}P^T\delta DP, \nabla^2\phi(\tau(D)) \cdot (-\frac{1}{2}(P^T\delta DP)) \rangle_F \\ &\quad + \text{h.o.t.} \\ &= \phi(\tau(D)) - \frac{1}{2}\langle P\nabla\phi(\tau(D))P^T, \delta D \rangle_F \\ &\quad + \frac{1}{2}\langle \delta D, \frac{1}{4}P(\nabla^2\phi(\tau(D)) \cdot P^T\delta DP)P^T \rangle_F + \text{h.o.t.},\end{aligned}$$

from which we see that

$$\nabla\Phi(D) = -\frac{1}{2} P \nabla\phi(\tau(D)) P^T, \quad (17)$$

$$\nabla^2\Phi(D) \cdot \delta D = \frac{1}{4} P (\nabla^2\phi(\tau(D)) \cdot P^T \delta D P) P^T. \quad (18)$$

Combining (15) and (17), we obtain

$$\nabla\Phi(D) = -\lambda P v v^T P^T.$$

We can simplify this further. Suppose first that $\lambda \neq 0$. Note that

$$-\frac{1}{2} P^T D P e = 0,$$

so e is an eigenvalue of $-\frac{1}{2} P^T D P$. If $\lambda \neq 0$, then, because eigenvectors associated with distinct eigenvalues are orthogonal, we have

$$P v = \left(I - \frac{1}{n} e e^T \right) v = v$$

and thus

$$\nabla\Phi(D) = -\lambda v v^T.$$

On the other hand, if $\lambda = 0$, then, because we are assuming λ is simple, it must be the case that v and e are collinear, so

$$\nabla\Phi(D) = 0 = -\lambda v v^T.$$

Thus, we always have

$$\nabla\Phi(D) = -\lambda v v^T. \quad (19)$$

Meanwhile, let

$$w = (A - \lambda I)^+ P^T \delta D P v,$$

where $A = \tau(D)$. Then (16) and (18) yield

$$\nabla^2\Phi(D) \cdot \delta D = \frac{1}{2} P (v v^T P^T \delta D P v v^T + \lambda (w v^T + v w^T)) P.$$

Again, first consider the case when $\lambda \neq 0$. As before, $P v = v$, so

$$\nabla^2\Phi(D) \cdot \delta D = \frac{1}{2} (v v^T \delta D v v^T + \lambda (P w v^T + v w^T P)).$$

Now,

$$w = (A - \lambda I)^+ P^T \delta D P v = (A - \lambda I)^+ P^T \delta D v,$$

so

$$\begin{aligned} \nabla^2\Phi(D) \cdot \delta D = & \frac{1}{2} \left(v v^T \delta D v v^T \right. \\ & \left. + \lambda \left[P (A - \lambda I)^+ P^T \delta D v v^T + v v^T \delta D P (A - \lambda I)^+ P^T \right] \right). \quad (20) \end{aligned}$$

On the other hand, if $\lambda = 0$, then, because v and e are collinear, we immediately see that

$$\nabla^2 \Phi(D) \cdot \delta D = 0$$

for all δD . That is, the Hessian vanishes.

4.3. The derivatives of the strain criterion

Finally, we come to the derivatives of

$$F(D) = F_3 \circ \tau(D) = \left\| -\frac{1}{2} P^T D P \right\|_F^2 - \sum_{i=1}^3 \max(0, \lambda_i)^2,$$

where λ_i are the eigenvalues of $\tau(D) = -\frac{1}{2} P^T D P$, ordered so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. We assume that F is differentiable at D . From the discussion in Section 3, we know that F is continuously differentiable provided $\lambda_1, \lambda_2, \lambda_3 \neq 0$ and, if $\lambda_3 > 0$, we also have $\lambda_4 \neq \lambda_3$.

For convenience, let

$$\bar{\lambda}_i = \max(0, \lambda_i), \quad \sigma(\lambda_i) = \begin{cases} +1 & \text{if } \lambda_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

If $\lambda_1, \lambda_2, \lambda_3$ are all simple, then we see that the gradient of F is given by

$$\nabla F(D) = \frac{1}{2} P^T D P + \sum_{i=1}^3 \bar{\lambda}_i v_i v_i^T, \quad (21)$$

where v_i are the eigendirections of $\tau(D)$ corresponding to λ_i .

Because we assume $\lambda_1, \lambda_2, \lambda_3 \neq 0$, simplification (20) is valid. The action of the Hessian is given by

$$\begin{aligned} \nabla^2 F(D) \cdot \delta D &= \frac{1}{2} P^T \delta D P - \frac{1}{2} \sum_{i=1}^3 \sigma(\lambda_i) v_i v_i^T \delta D v_i v_i^T \\ &\quad + \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i \left(P(A - \lambda_i I)^+ P^T \delta D P v_i v_i^T \right. \\ &\quad \left. + v_i v_i^T P^T \delta D P (A - \lambda_i I)^+ P^T \right). \end{aligned}$$

Moreover, if F is twice-differentiable at D , then we know that $\lambda_1, \lambda_2, \lambda_3 \neq 0$, and $e \perp \{v_1, v_2, v_3\}$, because e is a zero eigenvector (null vector) of $\tau(D)$. Thus we arrive at

$$\begin{aligned} \nabla^2 F(D) \cdot \delta D &= \frac{1}{2} P^T \delta D P - \frac{1}{2} \sum_{i=1}^3 \sigma(\lambda_i) v_i v_i^T \delta D v_i v_i^T \\ &\quad + \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i \left(P(A - \lambda_i I)^+ P^T \delta D v_i v_i^T \right. \\ &\quad \left. + v_i v_i^T \delta D P (A - \lambda_i I)^+ P^T \right). \quad (22) \end{aligned}$$

The preceding expressions for the gradient and Hessian have been derived under the assumption that $\lambda_1, \lambda_2, \lambda_3$ are simple eigenvalues. However, these expressions can be extended by continuity if any of the three largest eigenvalues coalesce. We omit the proof for reasons of length.

Many nonlinear programming algorithms require the Hessian matrix of the objective function in order to exploit second-order information. Because the Hessian matrix of F contains $O(n^4)$ entries, minimizing $F = F_3 \circ \tau$ by such algorithms is impractical. Having a formula for the Hessian-vector product makes it possible to minimize $F = F_3 \circ \tau$ by algorithms that exploit second-order information without requiring assembly of the entire Hessian matrix.

5. The spectrum of the Hessian

We can completely describe the spectrum of the Hessian of $F = F_3 \circ \tau$. This is rarely possible in nonlinear optimization; that it is possible here is noteworthy in its own right. More importantly, spectral information about the Hessian is of practical value in developing algorithms for solving optimization problems that involve F .

Most iterative methods for nonlinear optimization require minimizing successive quadratic Taylor's series approximations of the objective function. Quadratic functions of a moderate number of variables can be minimized efficiently by the techniques of direct linear algebra; however, because the problems that interest us are so large, we must resort to nonlinear programming methods based on iterative linear algebra techniques, e.g., conjugate gradient methods. Examples of such methods include the algorithms of [Moré and Toraldo \(1991\)](#) and [Moré and Lin \(1999\)](#).

If one knows the complete set of eigenvalues and eigenvectors for the Hessian, then one can re-scale the variables in ways that accelerate the linear algebra computations. This is called preconditioning. Knowledge of the spectrum is also useful when searching for directions of descent. If the Hessian is not positive definite, then one may want to modify it (as little as possible) so that it is. Knowledge of the Hessian's negative eigenvalues and associated eigenvectors allows precise modification of the Hessian to achieve positive definiteness.

Let $\{v_1, \dots, v_n\}$ be an orthonormal set of eigenvectors for $A = \tau(D)$. Define $\tilde{e} = e/\sqrt{n}$. Without loss of generality we may assume that \tilde{e} is one of the v_i , because e is an eigenvector (a null vector) of A .

Define

$$V_{jk} = v_j v_k^T + v_k v_j^T, \quad j \leq k. \quad (23)$$

These form an orthogonal basis for the symmetric matrices. (The scaling is not quite right for orthonormality.)

Theorem 5.1. *The matrices V_{jk} are the eigenvectors of $\nabla^2 f(D)$.*

The proof is a tedious calculation for each term in (22).

5.1. The first term in the Hessian

The first term in (22) is $\frac{1}{2} P^T V_{jk} P$. We have

$$\begin{aligned} P^T V_{jk} P &= \left(I - \frac{1}{n} e e^T \right) V_{jk} \left(I - \frac{1}{n} e e^T \right) \\ &= V_{jk} - \frac{1}{n} e e^T V_{jk} - \frac{1}{n} V_{jk} e e^T + \frac{1}{n^2} (e^T V_{jk} e) e e^T. \end{aligned}$$

There are two cases to consider.

5.1.1. Case (a): $\tilde{e} \in \{v_j, v_k\}$

In this case,

$$\frac{1}{2} P^T V_{jk} P = 0, \quad (24)$$

because $P^T v_j v_k^T P = 0$.

5.1.2. Case (b): $\tilde{e} \notin \{v_j, v_k\}$

In this case, $V_{jk} e = 0$, so

$$\frac{1}{2} P^T V_{jk} P = \frac{1}{2} V_{jk}. \quad (25)$$

5.2. The second term in the Hessian

The second term in (22) is $\frac{1}{2} \sum_{i=1}^3 \sigma(\lambda_i) v_i v_i^T V_{jk} v_i v_i^T$. We have

$$v_i^T V_{jk} v_i = v_i^T (v_j v_k^T + v_k v_j^T) v_i = \delta_{ij} \delta_{ki} + \delta_{ki} \delta_{ij} = 2\delta_{ij} \delta_{ik},$$

where δ_{ij} is now the Kronecker delta: $\delta_{ij} = 1$ if $i = j$; otherwise, $\delta_{ij} = 0$. Hence,

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^3 \sigma(\lambda_i) v_i v_i^T \delta D v_i v_i^T &= \frac{1}{2} \sum_{i=1}^3 \sigma(\lambda_i) (2\delta_{ij} \delta_{ik} v_i v_i^T) \\ &= \begin{cases} \sigma(\lambda_j) \frac{1}{2} V_{jk} & \text{if } j = k = i \text{ for some } i \in \{1, 2, 3\} \\ 0 & \text{otherwise} \end{cases} \quad (26) \end{aligned}$$

5.3. The third term in the Hessian

The third term in (22) is

$$\frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i \left(P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T + v_i v_i^T V_{jk} P(A - \lambda_i I)^+ P^T \right).$$

We begin by noting the following. If $\tilde{e} \notin \{v_j, v_k\}$, then e is orthogonal to v_j, v_k . As a consequence, $P v_j = v_j$, $P v_k = v_k$, and $P^T V_{jk} = V_{jk}$. The third term then

simplifies as follows:

$$\begin{aligned}
& \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T \\
&= \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ V_{jk} v_i v_i^T \\
&= \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ (v_j v_k^T + v_k v_j^T) v_i v_i^T \\
&= \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ (v_j \delta_{ik} + v_k \delta_{ij}) v_i^T \\
&= \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P [(\lambda_j - \lambda_i)^+ v_j \delta_{ik} + (\lambda_k - \lambda_i)^+ v_k \delta_{ij}] v_i^T.
\end{aligned}$$

(For a scalar α , by α^+ we mean $\alpha^+ = 1/\alpha$ if $\alpha \neq 0$, and 0 otherwise.) Finally, we arrive at

$$\begin{aligned}
& \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T \\
&= \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i [(\lambda_j - \lambda_i)^+ v_j \delta_{ik} + (\lambda_k - \lambda_i)^+ v_k \delta_{ij}] v_i^T. \tag{27}
\end{aligned}$$

This holds provided $\tilde{e} \notin \{v_j, v_k\}$.

5.3.1. Case (a): $\tilde{e} \in \{v_j, v_k\}$

In this case, we know that $P^T V_{jk} P = 0$. Because $P^T V_{jk} v_i v_i = P^T V_{jk} P v_i v_i$, the third term in the Hessian vanishes:

$$\frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i (P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T + v_i v_i^T V_{jk} P(A - \lambda_i I)^+ P^T) = 0. \tag{28}$$

5.3.2. Case (b): $\tilde{e} \notin \{v_j, v_k\}$, $j, k \notin \{1, 2, 3\}$

If $j, k \notin \{1, 2, 3\}$, then $\delta_{ij} = \delta_{ik} = 0$ for $i = 1, 2, 3$, and (27) yields

$$\frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i \left(P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T + v_i v_i^T V_{jk} P(A - \lambda_i I)^+ P^T \right) = 0. \tag{29}$$

5.3.3. Case (c): $\tilde{e} \notin \{v_j, v_k\}$, $j \in \{1, 2, 3\}$, $k \notin \{1, 2, 3\}$

Because $k \notin \{1, 2, 3\}$, we know that $\delta_{ik} = 0$ for $i = 1, 2, 3$. Thus, (27) yields

$$\frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T = \frac{1}{2} \bar{\lambda}_j (\lambda_k - \lambda_j)^+ v_k v_j^T.$$

In this case the third term of the Hessian is

$$\begin{aligned}
 & \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i \left(P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T + v_i v_i^T V_{jk} P(A - \lambda_i I)^+ P^T \right) \\
 &= \frac{1}{2} \left[\bar{\lambda}_j (\lambda_k - \lambda_j)^+ v_k v_j^T + \frac{1}{2} \bar{\lambda}_j (\lambda_k - \lambda_j)^+ v_j v_k^T \right] \\
 &= \frac{1}{2} \bar{\lambda}_j (\lambda_k - \lambda_j) V_{jk}. \tag{30}
 \end{aligned}$$

5.3.4. *Case (d):* $\tilde{e} \notin \{v_j, v_k\}$, $j \notin \{1, 2, 3\}$, $k \in \{1, 2, 3\}$

This is not actually a case we need consider, as we assume $j \leq k$ in the definition of V_{jk} in (23). We mention it only to reassure ourselves that we have not overlooked a possibility.

5.3.5. *Case (e):* $\tilde{e} \notin \{v_j, v_k\}$, $j, k \in \{1, 2, 3\}$

This time, (27) tells us that

$$\begin{aligned}
 & \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T \\
 &= \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i [(\lambda_j - \lambda_i)^+ v_j \delta_{ik} + (\lambda_k - \lambda_i)^+ v_k \delta_{ij}] v_i^T \\
 &= \frac{1}{2} \left[\bar{\lambda}_k (\lambda_j - \lambda_k)^+ v_j v_k^T + \bar{\lambda}_j (\lambda_k - \lambda_j)^+ v_k v_j^T \right].
 \end{aligned}$$

Thus, for the third term of the Hessian we obtain

$$\begin{aligned}
 & \frac{1}{2} \sum_{i=1}^3 \bar{\lambda}_i \left(P(A - \lambda_i I)^+ P^T V_{jk} v_i v_i^T + v_i v_i^T V_{jk} P(A - \lambda_i I)^+ P^T \right) \\
 &= \frac{1}{2} \left[\bar{\lambda}_k (\lambda_j - \lambda_k)^+ v_j v_k^T + \bar{\lambda}_j (\lambda_k - \lambda_j)^+ v_k v_j^T \right] \\
 &\quad + \frac{1}{2} \left[\bar{\lambda}_j (\lambda_k - \lambda_j)^+ v_k v_j^T + \bar{\lambda}_k (\lambda_j - \lambda_k)^+ v_j v_k^T \right] \\
 &= \frac{1}{2} \left[\bar{\lambda}_j (\lambda_k - \lambda_j)^+ + \bar{\lambda}_k (\lambda_j - \lambda_k)^+ \right] V_{jk}. \tag{31}
 \end{aligned}$$

5.4. Putting it all together

Now we can write down the spectrum of the Hessian. Denoting the Hessian by H , we have

- (1) $\tilde{e} \in \{v_j, v_k\}$: $HV_{jk} = 0$, from (24), (26), and (28), and the fact that $\tilde{e} \notin \{v_1, v_2, v_3\}$.
- (2) $\tilde{e} \notin \{v_j, v_k\}$, $j, k \notin \{1, 2, 3\}$: $HV_{jk} = \frac{1}{2} V_{jk}$, from (25), (26), and (29).
- (3) $\tilde{e} \notin \{v_j, v_k\}$, $j \in \{1, 2, 3\}$, $k \notin \{1, 2, 3\}$: From (25), (26), and (30),

$$HV_{jk} = \frac{1}{2} V_{jk} - 0 + \frac{1}{2} \bar{\lambda}_j (\lambda_k - \lambda_j)^+ V_{jk} = \frac{1}{2} \left[1 + \bar{\lambda}_j (\lambda_k - \lambda_j)^+ \right] V_{jk}.$$

Because $j \in \{1, 2, 3\}$ but $k \notin \{1, 2, 3\}$, we know that $\lambda_j \neq \lambda_k$; hence, the preceding can be written as

$$HV_{jk} = \frac{1}{2} \left[1 + \frac{\bar{\lambda}_j}{\lambda_k - \lambda_j} \right] V_{jk}.$$

(3') $\tilde{e} \notin \{v_j, v_k\}$, $j \notin \{1, 2, 3\}$, $k \in \{1, 2, 3\}$: As noted in Section 5.3, this is not a possibility as we assume $j \leq k$.

(4) $\tilde{e} \notin \{v_j, v_k\}$, $j, k \in \{1, 2, 3\}$, $j \neq k$: From (25), (26), and (31),

$$\begin{aligned} HV_{jk} &= \frac{1}{2} V_{jk} - 0 + \frac{1}{2} \left[\bar{\lambda}_j (\lambda_k - \lambda_j)^+ + \bar{\lambda}_k (\lambda_j - \lambda_k)^+ \right] V_{jk} \\ &= \frac{1}{2} \left[1 - (\bar{\lambda}_k - \bar{\lambda}_j) (\lambda_k - \lambda_j)^+ \right] V_{jk}. \end{aligned}$$

(5) $\tilde{e} \notin \{v_j, v_k\}$, $j, k \in \{1, 2, 3\}$, $j = k$: From (25), (26), and (31),

$$HV_{jk} = \frac{1}{2} V_{jk} - \frac{1}{2} V_{jk} + \frac{1}{2} \left[\bar{\lambda}_j (\lambda_k - \lambda_j)^+ + \bar{\lambda}_k (\lambda_j - \lambda_k)^+ \right] V_{jk} = 0.$$

To summarize, let A_{jk} be the eigenvalue associated with V_{jk} . Then the spectrum of the Hessian is $A_{jk} = 0$ if $\tilde{e} \in \{v_j, v_k\}$ and

$$A_{jk} = \begin{cases} \frac{1}{2} & \text{if } j, k \notin \{1, 2, 3\}, \\ \frac{1}{2} \left[1 + \frac{\bar{\lambda}_j}{\lambda_k - \lambda_j} \right] & \text{if } j \in \{1, 2, 3\}, k \notin \{1, 2, 3\}, \\ \frac{1}{2} \left[1 - (\bar{\lambda}_k - \bar{\lambda}_j) (\lambda_k - \lambda_j)^+ \right] & \text{if } j, k \in \{1, 2, 3\}, j \neq k, \\ 0 & \text{if } j, k \in \{1, 2, 3\}, j = k \end{cases}$$

if $\tilde{e} \notin \{v_j, v_k\}$. Further simplifications are possible when λ_j and λ_k are positive.

5.5. Interpretation of the spectrum

Some observations concerning the spectrum are in order. Let Δ be a matrix of squared dissimilarities. Most significantly, the calculation of the spectrum of the strain Hessian shows that the nonconvexity of F , and thus the difficulty in its minimization, is related to the realizability of Δ as dissimilarities of a set of points in \mathbb{R}^p , for some $p > 3$.

To see this, suppose first that $j \in \{1, 2, 3\}$ and $\lambda_j > 0$. Then, if λ_k is another eigenvalue of $\tau(\Delta)$ and $\lambda_k \leq 0$, the associated eigenvalue A_{jk} of the Hessian lies in the interval $[0, 1]$. If $\lambda_k > 0$, on the other hand, then the associated eigenvalue A_{jk} of the Hessian will be negative, indicating local nonconvexity of F .

Suppose next that the eigenvalues λ_k are all nonnegative, and that $\lambda_1, \dots, \lambda_p$ are positive. Then Δ corresponds to a set of points that can be embedded in \mathbb{R}^p . However, if $p > 3$, then the eigenvalues $\lambda_4, \dots, \lambda_p$ of $\tau(\Delta)$ lead to the negative eigenvalues of the Hessian of F , and thus nonconvexity of the strain criterion.

On the other hand, eigenvalues λ_k of $\tau(\Delta)$ that are negative have no quasi-physical interpretation. Observe that they correspond to nonnegative eigenvalues of the Hessian of F , or directions of convexity (positive curvature). These, in principle, are more easily handled by an optimization algorithm.

Note, too, that the large eigenvalues of the Hessian arise if the eigenvalue λ_3 of $\tau(\Delta)$ is positive and does not lie far from λ_4 . This reflects the nonlinearity associated with coalescing eigenvalues (e.g., loss of regularity).

6. $F_3 \circ \tau$ as a function of hollow symmetric matrices

So far, we have studied $F = F_3 \circ \tau$ as a function of general symmetric matrices. However, a matrix of squared dissimilarities is also hollow, i.e., its diagonal entries vanish. Because the representation of the first derivative of F given in (21) may not be hollow, it is not an appropriate representer on the space of hollow symmetric matrices. A similar comment holds for the action of the Hessian described in (22): even if δD has a zero diagonal, $\nabla^2 F \cdot \delta D$ need not. Accordingly, we now study F as a function of hollow symmetric matrices.

6.1. The projected Hessian

One way to account for the restriction to hollow matrices is to consider the projected gradient and projected Hessian. Let d_{ij} denote the entries of a symmetric matrix D . The projection operator ζ taking D to the space of hollow matrices is just

$$\zeta : D \mapsto \zeta(D), \text{ where } (\zeta(D))_{ij} = \begin{cases} 0 & \text{if } i = j, \\ d_{ij} & \text{otherwise.} \end{cases}$$

Let \tilde{F} denote F restricted to the hollow matrices. Then the gradient of \tilde{F} is simply the projected gradient: $\nabla \tilde{F}(\Delta) = \zeta^* \nabla F(\Delta)$. Because ζ is self-adjoint, this reduces to $\nabla \tilde{F}(\Delta) = \zeta \nabla f(\Delta)$. Similarly, the Hessian of \tilde{F} is the projected Hessian: $\nabla^2 \tilde{F}(\Delta) = \zeta^* \nabla^2 F(\Delta) \zeta$.

Unfortunately, with the introduction of the projection operator we lose the nice characterization of the spectrum of the Hessian given in the preceding section. Nevertheless, we still know something about the projected Hessian because it is congruent to the Hessian on the space of all symmetric matrices D . For instance, we know the inertia of the projected Hessian, and can bound the size of the extreme eigenvalues of the projected Hessian.

6.2. The reduced Hessian

Alternatively, we can view the hollow matrices as the range of the space of symmetric matrices under the mapping we define as follows. Given a matrix D , let $\text{diag}(D)$ denote the vector in \mathbb{R}^n whose components are the diagonal entries of D . Define ζ by

$$\zeta : D \mapsto D - \frac{1}{2} \left[e \text{diag}(D)^T + \text{diag}(D) e^T \right].$$

Notice that $\zeta(D) = -\kappa(D)$, where [Critchley's \(1988\)](#) κ and τ are mutually inverse on the appropriate subspaces. If D is symmetric, then $\zeta(D)$ is symmetric and hollow. Moreover, ζ maps the symmetric matrices onto the hollow symmetric matrices. Thus we can view the problem of minimizing $F = F_3 \circ \tau$ on the space of the hollow symmetric matrices as the problem of minimizing $\hat{F} = F_3 \circ \tau \circ \zeta$ on the space of symmetric matrices.

The gradient of \hat{F} is then the *reduced gradient*,

$$\nabla \hat{F}(D) = \zeta^* \nabla F(D)$$

and the Hessian of \hat{F} is the *reduced Hessian*,

$$\nabla^2 \hat{F}(D) = \zeta^* \nabla^2 F(D) \zeta.$$

(Of course, in this context there is no true variable reduction because the domain of ζ is of higher dimension than its range.)

The adjoint of ζ is easily computed: if C and D are symmetric, then

$$\begin{aligned} \langle C, \zeta(D) \rangle_F &= \text{trace} \left(C \left(D - \frac{1}{2} \left[e \text{diag}(D)^T + \text{diag}(D) e^T \right] \right) \right) \\ &= \text{trace}(CD) - eC \text{diag}(D)^T \\ &= \langle D, C - \text{Diag}(Ce) \rangle_F, \end{aligned}$$

where we have applied (2). Thus,

$$\zeta^* : C \mapsto C - \text{Diag}(Ce).$$

Next we discuss the spectrum of the reduced Hessian. For the purposes of the discussion that follows, let θ be the mapping

$$\theta : D \mapsto \frac{1}{2} \left[e \text{diag}(D)^T + \text{diag}(D) e^T \right],$$

so $\zeta = I - \theta$.

Recall the following: if we are at a point where F (and thus \hat{F}) is twice-differentiable, then $\lambda_1, \lambda_2, \lambda_3 \neq 0$. Because e is a zero eigenvector (null vector) of $\tau(D)$, $e \perp \{v_1, v_2, v_3\}$.

Let D be any symmetric matrix, and let $d = \text{diag}(D)$. Write d in terms of the orthonormal set of eigenvectors v_i of $\tau(D)$:

$$d = \sum_{i=1}^n \alpha_i v_i.$$

Then

$$\theta(D) = \frac{1}{2} \sum_{i=1}^n \alpha_i \left[e v_i^T + v_i e^T \right].$$

Because $A_{jk} = 0$ if $\tilde{e} \in \{v_j, v_k\}$, $\theta(D)$ must be a null vector of $\nabla^2 F$. Because this is true for any D , we have $\nabla^2 F \theta = 0$ and $\theta^* \nabla^2 F = 0$. Thus, the reduced Hessian satisfies

$$\nabla^2 \hat{F} = \zeta^* \nabla^2 F \zeta = (I - \theta) \nabla^2 F (I - \theta) = \nabla^2 F$$

and the eigenstructure of the reduced Hessian is the same as the eigenstructure of the original Hessian!

The preceding suggests one way to use our knowledge of the spectrum of the Hessian of F . If we wish to minimize $F = F_3 \circ \tau$, then we may wish to parameterize the hollow symmetric matrices in terms of general symmetric matrices via ζ . This poses the optimization problem in terms of the symmetric matrices.

7. Concluding remarks

Classical MDS can be stated as the problem of finding the symmetric positive semidefinite matrix of rank $\leq p$ that is nearest $\tau(\Delta)$ is squared Frobenius distance, for Δ fixed. This problem has an explicit solution, and the minimum squared distance is sometimes called the strain criterion. We have been concerned with extensions of classical MDS in which Δ is free to vary. (For simplicity, we have emphasized the important case of $p = 3$.) The resulting nonlinear optimization problems do not have explicit solutions; iterative methods are required.

A great deal is known about the geometry of the closed cone of symmetric positive semidefinite matrices of rank $\leq p$, but it is not clear how to exploit this knowledge in an optimization algorithm. Traditional nonlinear programming assumes a problem that has been formulated using (preferably linear) equality and inequality constraints. Accordingly, Trosset proposed using the strain criterion as an objective function, thereby eliminating a nonlinear constraint that confounds traditional optimization algorithms. The geometry of that constraint is preserved, encoded in the objective function. The challenge addressed in the preceding sections is how to extract that information in a form that can be used by traditional optimization algorithms.

The preceding sections have provided a complete characterization of the Hessian of the strain criterion. The strain criterion is unusual insofar as it is a highly nontrivial nonlinear function, yet its Hessian-vector products can be computed in closed form, as can the eigenvalues and eigenvectors of the Hessian. We expect that these properties can be exploited, leading to more efficient algorithms for solving the optimization problems that motivated the present study. For example, Trosset (2002) used a bound-constrained quasi-Newton algorithm with a limited memory BFGS updating formula (LM-BFGS) that constructs Hessian approximations from gradients computed for several previous iterations. For unconstrained optimization, Nash and Nocedal (1991) compared LM-BFGS and a truncated Newton strategy. The latter is based on the successive approximate minimization, via conjugate gradients, of a quadratic model of the nonlinear objective—an approach that is similar to the algorithms that we are developing for constrained minimization of the strain criterion. Although general comparisons are difficult, Nash and Nocedal observed that, for certain conditions that obtain in the strain objective (e.g., a mildly nonlinear objective, a Hessian with some zero eigenvalues), truncated Newton was generally more efficient than LM-BFGS. Truncated Newton is even more attractive when Hessian-vector products can be computed efficiently, as can be done for the strain criterion using (22).

Finally, Trosset (2002) observed that working directly with squared dissimilarities, rather than parameterizing distance matrices by the Cartesian coordinates of n points in \mathbb{R}^p , appears to result in more tractable optimization problems with fewer nonglobal minimizers. Intuitively, it would seem that allowing each squared dissimilarity to vary independently of the others has the positive effect of decoupling the complicated interactions that result from varying individual coordinates. Our analysis of the Hessian of the strain criterion clarifies the nature of its nonconvexity, yielding new insight. Specifically, the nonconvexity of the strain criterion near Δ is related to the possibility of embedding Δ in some \mathbb{R}^q , where $q > p$. This insight contradicts, at least superficially, popular speculation that one would do well

to address problems in \mathbb{R}^p by working in \mathbb{R}^q . The full significance of this insight is not yet clear to us; it will be investigated in future work.

References

- Borg, I., Groenen, P., 1997. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York.
- Cox, T.F., Cox, M.A.A., 1994. *Multidimensional Scaling*. Chapman & Hall, London.
- Critchley, F., 1988. On certain linear mappings between inner-product and squared-distance matrices. *Linear Algebra Appl.* 105, 91–107.
- de Leeuw, J., Heiser, W., 1982. Theory of multidimensional scaling. Krishnaiah, P.R., Kanai, I.N. (Eds.), *Handbook of Statistics*, vol. 2. North-Holland, Amsterdam, pp. 285–316 (Chapter 13).
- Everitt, B.S., Dunn, G., 1991. *Applied Multivariate Data Analysis*. Edward Arnold, London.
- Everitt, B.S., Rabe-Hesketh, S., 1997. *The Analysis of Proximity Data*. Edward Arnold, London.
- Gower, J.C., 1966. Some distance properties of latent root and vector methods in multivariate analysis. *Biometrika* 53, 315–328.
- Gower, J.C., Groenen, P.J.F., 1991. Applications of the modified Leverrier–Faddeev algorithm for the construction of explicit matrix spectral decompositions and inverses. *Utilitas Math.* 40, 51–64.
- Krzyszowski, W.J., Marriott, F.H.C., 1994. *Multivariate Analysis Part 1 Distributions, Ordination and Inference*. Edward Arnold, London.
- Lewis, A.S., 1996. Derivatives of spectral functions. *Math. Oper. Res.* 21, 576–588.
- Lewis, A.S., Sendov, H.S., 2001. Twice differentiable spectral functions. *SIAM J. Matrix Anal. Appl.* 23, 368–386.
- Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. *Multivariate Analysis*. Academic Press, Orlando.
- Moré, J.J., Lin, C.-H., 1999. Newton’s method for large bound-constrained optimization problems. *SIAM J. Optim.* 9, 1100–1127.
- Moré, J.J., Toraldo, G., 1991. On the solution of large quadratic programming problems with bound constraints. *SIAM J. Optim.* 1, 93–113.
- Nash, S.G., Nocedal, J., 1991. A numerical study of the limited memory BFGS method and the truncated-Newton method for large scale optimization. *SIAM J. Optim.* 1, 358–372.
- Parks, T.A., 1985. Reducible nonlinear programming problems. Technical Report 85-8, Department of Mathematical Sciences, Rice University, Houston, TX, USA.
- Seber, G.A.F., 1984. *Multivariate Observations*. Wiley, New York.
- Sibson, R., 1979. Studies in the robustness of multidimensional scaling: perturbational analysis of classical scaling. *J. Roy. Statist. Soc. Ser. B* 41, 217–229.
- Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17, 401–419.
- Trosset, M.W., 1997a. Computing distances between convex sets and subsets of the positive semidefinite matrices. Technical Report 97-3, Department of Computational & Applied Mathematics, Rice University, Houston, TX, USA.
- Trosset, M.W., 1997b. Numerical algorithms for multidimensional scaling. In: Klar, R., Opitz, P. (Eds.), *Classification and Knowledge Organization*. Springer, Berlin, pp. 80–92.
- Trosset, M.W., 1998a. Applications of multidimensional scaling to molecular conformation. *Comput. Sci. Statist.* 29, 148–152.
- Trosset, M.W., 1998b. A new formulation of the nonmetric STRAIN problem in multidimensional scaling. *J. Classification* 15, 15–35.
- Trosset, M.W., 2000. Distance matrix completion by numerical optimization. *Comput. Optim. Appl.* 17, 11–22.
- Trosset, M.W., 2002. Extensions of classical multidimensional scaling via variable reduction. *Comput. Statist.* 17, 147–163.
- Wilkinson, J.H., 1965. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, England.