Computing Distances Between Convex Sets and Subsets of the Positive Semidefinite Matrices

Michael W. Trosset*

June 7, 1997

Abstract
We describe an important class of semidefinite programming problems that has received scant attention in the optimization community. These problems are derived from considerations in distance geometry and multidimensional scaling and therefore arise in a variety of disciplines, e.g., computational chemistry and psychometrics. In most applications, the feasible positive semidefinite matrices are restricted in rank, so that recent interior-point methods for semidefinite programming do not apply. We establish some theory for these problems and discuss what remains to be accomplished.

Key words: Distance geometry, multidimensional scaling, semidefinite programming.

Contents
1 Introduction .................................................. 2
2 Projection into Subsets of $\Omega_n$ ....................... 4
3 Reducible Programming Formulations ................. 5
  3.1 Variable Alternation .................................... 5
  3.2 Variable Reduction ..................................... 6
4 Optimization by Variable Alternation ................. 7
5 Optimization by Variable Reduction ................... 9
6 Discussion .................................................. 13

*Adjunct Associate Professor, Department of Computational & Applied Mathematics—MS 134, Rice University, 6100 Main Street Houston, TX 77005-1892 (email: trosset@caam.rice.edu).
1 Introduction

This report introduces a new class of optimization problems. These problems are derived from considerations in distance geometry and multidimensional scaling and are therefore of great significance in such diverse disciplines as computational chemistry and psychometrics. Better algorithms for solving these problems would be of enormous value.

In this report we are concerned with mathematical structure and will make only passing reference to specific applications. We begin with some definitions:

**Definition 1** A hollow matrix is a square matrix whose diagonal elements vanish.

**Definition 2** A dissimilarity matrix is a symmetric hollow matrix with nonnegative elements.

We will denote dissimilarity matrices by $\Delta = [\delta_{ij}]$. The $\delta_{ij}$ are called dissimilarities.

**Definition 3** A $p$-dimensional Euclidean distance matrix is an $n \times n$ matrix $D = [d_{ij}]$ for which there exist $x_1, \ldots, x_n \in \mathbb{R}^p$ such that $d_{ij} = \|x_i - x_j\|$.

We will denote the set of $n \times n$ $p$-dimensional distance matrices by $\mathcal{D}_n(p)$. Given a set of points $x_1, \ldots, x_n \in \mathbb{R}^p$, we store the coordinates of $x_i$ in row $i$ of the $n \times p$ configuration matrix $X$ and denote the matrix of interpoint distances by $D(X)$.

It is obvious that a distance matrix is necessarily a dissimilarity matrix. Determining whether or not a specified dissimilarity matrix is a distance matrix is a famous problem in classical distance geometry. We state the standard solution of this problem, implicit in Torgerson’s [24] formulation of multidimensional scaling and demonstrated by Gower [9]. The standard solution is a trivial modification of the solution independently discovered by Schoenberg [20] and by Young and Householder [31]. Its statement requires some additional definitions:

**Definition 4** Let $A = [a_{ij}]$ and $B = [b_{ij}]$ denote $m \times n$ matrices. The Hadamard product of $A$ and $B$ is $A \ast B = [a_{ij}b_{ij}]$.

Thus, if $\Delta$ is a matrix of dissimilarities, then $\Delta \ast \Delta = [\delta^2_{ij}]$ is a matrix of squared dissimilarities. Notice that $\Delta \ast \Delta$ is itself a dissimilarity matrix.

**Definition 5** The double-centering operator $\tau$ is a linear mapping on square matrices. If $A = [a_{ij}]$ is an $n \times n$ matrix, then $B = [b_{ij}] = \tau(A)$ is the $n \times n$ matrix defined by

$$b_{ij} = -\frac{1}{n} \left( a_{ij} - \frac{1}{n} \sum_{j=1}^{n} a_{ij} - \frac{1}{n} \sum_{i=1}^{n} a_{ij} + \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij} \right).$$

Let $I_n$ denote the $n \times n$ identity matrix and let $1_n \in \mathbb{R}^n$ denote the vector whose elements are 1. Then double centering can also be represented by the matrix equation

$$\tau(A) = -\frac{1}{2} \left( I_n - \frac{1}{n} 1_n 1_n' \right) A \left( I_n - \frac{1}{n} 1_n 1_n' \right).$$  \hspace{1cm} (1)

Notice that $\tau(A)$ is symmetric if $A$ is symmetric. A detailed study of $\tau$ and related mappings was made by Critchley [3].

Let $\Omega_n$ denote the set of symmetric positive semidefinite $n \times n$ matrices. Let $\Omega_n(p)$ denote the matrices in $\Omega_n$ whose rank is no greater than $p$. We now state the embedding theorem on which our work is based:
Theorem 1 Let \( \Delta \) be an \( n \times n \) dissimilarity matrix. Then \( \Delta \in \mathcal{D}_n(p) \) if and only if \( \tau(\Delta * \Delta) \in \Omega_n(p) \). Furthermore, if the \( n \times p \) matrix \( X \) is such that \( X'X = \tau(\Delta * \Delta) \), then \( D(X) = \Delta \).

Of course, if \( B \in \Omega_n(p) \), then it is easy to use the spectral decomposition of \( B \) to construct an \( n \times p \) matrix \( X \) such that \( X'X = B \).

Given a set of dissimilarity matrices and a target dimension \( p \), distance geometry and multidimensional scaling are concerned with finding a configuration of points in \( \mathbb{R}^p \) whose interpoint distance matrix matches or approximates one of the specified dissimilarity matrices. These concerns can be formulated as optimization problems in which the objective functions measure the discrepancy between a dissimilarity matrix and a distance matrix. In metric multidimensional scaling, a single dissimilarity matrix is specified. In applications of distance geometry to molecular conformation, the set of dissimilarity matrices is specified by bound constraints on the dissimilarities and the target dimension is necessarily \( p = 3 \). In nonmetric multidimensional scaling, the set of dissimilarity matrices is specified by order constraints on the dissimilarities. A survey of several common formulations was made by Troset [28].

Because of Theorem 1, a natural way to measure the discrepancy between a dissimilarity matrix \( \Delta \) and a distance matrix \( D \) is

\[
\|\tau(\Delta * \Delta) - \tau(D * D)\|_F^2,
\]

where \( \| \cdot \|_F \) denotes the Frobenius norm. The constraints typically imposed on the dissimilarities are such that the constraints on \( \Delta * \Delta \) are closed and convex. If \( D \in \mathcal{D}_n(p) \), then \( \tau(D * D) \in \Omega_n(p) \subseteq \Omega_n \).

Thus, the optimization problems that we will consider are of the general form

\[
\begin{align*}
\text{minimize} & \quad \|\tau(\Delta) - B\|_F^2, \\
\text{subject to} & \quad \Delta \in \mathcal{C}_n, \\
& \quad B \in \mathcal{B}_n,
\end{align*}
\]

where \( \mathcal{C}_n \) is a closed convex set of dissimilarity matrices and \( \mathcal{B}_n \) is a closed subset of \( \Omega_n \). Because \( \tau \) is a linear transformation, \( \tau(\mathcal{C}_n) \) is a closed convex set; thus, Problem (2) is the problem of minimizing the distance between a (certain type of) closed convex set of symmetric matrices and a closed subset of the symmetric positive semidefinite matrices.

The set \( \Omega_n \) is closed and convex. Hence, if \( \mathcal{B}_n = \Omega_n \), then Problem (2) is convex. Unfortunately, we are aware of no applications for which \( \mathcal{B}_n = \Omega_n \). (For example, we have already noted that applications to computational chemistry require \( \mathcal{B}_n = \Omega_n(3) \).) Accordingly, the remainder of this report develops methods that can be applied to Problem (2) without assuming convexity.

In Section 2 we introduce the subsets \( \mathcal{B}_n \subseteq \Omega_n \) considered in this report. These subsets have a special structure, but are more general than the subsets \( \Omega_n(p) \). We restrict attention to the case of a single dissimilarity matrix, i.e. \( \mathcal{C}_n = \{ \Delta \} \), and derive an explicit global solution of Problem (2). Our result generalizes a characterization of classical multidimensional scaling due to Gower [9] and Mardia [13].

In Section 3 we restore the weaker assumption that \( \mathcal{C}_n \) is a closed convex set of dissimilarity matrices. We discuss the general structure of Problem (2) and describe two general optimization strategies, variable alternation and variable reduction, for exploiting that structure.

In Sections 4 and 5 we establish some properties of the variable alternation and variable reduction approaches to Problem (2). We emphasize the latter, for which the theory is more satisfying. Using this approach, we propose a gradient projection method for finding local solutions of Problem (2). Section 6 concludes with a discussion of the prospects for developing more efficient methods and for finding global solutions of Problem (2).
2 Projection into Subsets of $\Omega_n$

In this section we restrict attention to the case of a fixed dissimilarity matrix, i.e., $C_n = \{\Delta\}$. In this case, Problem (2) specializes to the problem of finding the matrix in $B_n \subseteq \Omega_n$ that is nearest the symmetric $n \times n$ matrix $T = \tau(\Delta)$. When $B_n = \Omega_n$, this is the problem of projecting $T$ into the closed convex cone of symmetric positive semidefinite matrices. In most applications, $B_n$ is not convex; often, however, it is still possible to find an element of $B_n$ that is nearest $T$.

To that end, let $K$ be any closed and convex subset of
\[ \{x \in \mathbb{R}^n : x_1 \geq \cdots \geq x_n \geq 0\} \quad (3) \]
and let $\Omega_n(K)$ denote the set of symmetric $n \times n$ matrices of the form $UTU'$, where $U$ is orthogonal and diag($\Gamma$) $\in K$. Notice that, if
\[ K = \{x \in \mathbb{R}^n : x_1 \geq \cdots \geq x_{p+1} = \cdots = 0\}, \]
then $\Omega_n(K) = \Omega_n(p)$. In this section we consider the following special case of Problem (2):
\begin{align*}
\text{minimize} & \quad \|T - B\|^2_F \\
\text{subject to} & \quad B \in \Omega_n(K). \quad (4)
\end{align*}

The following result generalizes Theorem 14.4.2 in [14]:

Theorem 2 Let $Q \Lambda Q'$ represent any spectral decomposition of $T$ for which the eigenvalues $\lambda = \text{diag}(\Lambda)$ satisfy $\lambda_1 \geq \cdots \geq \lambda_n$. Let $\pi$ denote projection into $K$, $\tilde{\lambda} = \pi \lambda$, and $\tilde{\Lambda} = \text{diag}(\tilde{\lambda})$. Then $B^* = Q \tilde{\Lambda} Q'$ is a global minimizer of Problem (4).

Proof: Given $B \in \Omega_n(K)$, write $B = UTU'$ and $R = Q'U$, whereby
\[ \|T - B\|^2_F = \|Q \Lambda Q' - UTU'\|^2_F = \|\Lambda - Q'UTU'Q\|^2_F = \|\Lambda - R \Gamma R'\|^2_F. \]

For any fixed $\Gamma = \text{diag}(\gamma)$,
\[ \|\Lambda - R \Gamma R'\|^2_F = \|\Lambda\|^2_F - 2 \langle \Lambda, R \Gamma R' \rangle_F + \|\Gamma\|^2_F \quad (5) \]
is minimized by choosing the orthogonal matrix $R = [r_{ij}]$ to maximize the inner product
\[ \langle \Lambda, R \Gamma R' \rangle_F = \text{tr} \langle \Lambda R \Gamma R' \rangle = \sum_{i=1}^n \lambda_i \left( \sum_{j=1}^n \gamma_j r_{ij}^2 \right) = \sum_{i=1}^n \alpha_i \lambda_i. \quad (6) \]

Notice that $\alpha_i \geq 0$ and
\[ \sum_{i=1}^n \alpha_i = \sum_{i=1}^n \sum_{j=1}^n \gamma_j r_{ij}^2 = \sum_{j=1}^n \gamma_j. \]

Because $\lambda_1 \geq \cdots \geq \lambda_n$, we maximize (6) by choosing $\alpha_1$ as large as possible. Because $\gamma_1 \geq \cdots \geq \gamma_n \geq 0$ and $R$ is orthogonal, this is accomplished by choosing $r_{11} = 1$, which forces $r_{12} = \cdots = r_{1n} = 0$. Similarly, we maximize $\sum_{i=2}^n \alpha_i \lambda_i$ by choosing $r_{22} = 1$, and so on. It follows that, for any $\Gamma$, (6) is maximized, hence (5) is minimized, by choosing $R = I$. We conclude that
\begin{align*}
\|\Lambda - B\|^2_F & \geq \|\Lambda - \Gamma\|^2_F \geq \|\lambda - \gamma\|^2_F \\
& \geq \|\lambda - \tilde{\lambda}\|^2_F = \|\Lambda - \tilde{\Lambda}\|^2_F = \|Q(\Lambda - \tilde{\Lambda}Q')\|^2_F = \|Q\Lambda Q' - Q\tilde{\Lambda}Q'\|^2_F \\
& = \|T - B^*\|^2_F,
\end{align*}

4
as claimed. □

We conclude this section by recording several expressions for the global minimum of Problem (4). For reasons that will become apparent in Section 5, we represent the minimum as a function, $F_K \circ \tau$, of $\Delta$. Because $\tau$ is a linear function of $\Delta$ and $F_K(T)$ is the squared distance from $T$ to $\Omega_n(K)$, $F_K \circ \tau$ is a continuous function of $\Delta$. In general,

$$F_K \circ \tau(\Delta) = \|Q\Delta Q' - Q\tilde{\Delta}Q'\|_F^2 = \|Q(\Delta - \tilde{\Delta})Q'\|_F^2 = \text{tr} \left( (\Delta - \tilde{\Delta})^2 \right) = \sum_{i=1}^{n}(\lambda_i - \tilde{\lambda}_i)^2. \quad (7)$$

If $\Omega_n(K) = \Omega_n(p)$, then these expressions specialize to

$$F_p \circ \tau(\Delta) = \sum_{i=1}^{p}[\lambda_i - \max(\lambda_i, 0)]^2 + \sum_{i=p+1}^{n} \lambda_i^2 = \sum_{i=1}^{p} \phi(\lambda_i) + \sum_{i=p+1}^{n} \psi(\lambda_i), \quad (8)$$

where $\phi$ and $\psi$ denote the continuously differentiable functions $\phi(t) = t^2 I_{(-\infty,0]}(t)$ and $\psi(t) = t^2$.

### 3 Reducible Programming Formulations

Problem (2) has the general form

$$\begin{align*}
\text{minimize} & \quad f(a, b) \\
\text{subject to} & \quad a \in A, \ b \in B. 
\end{align*} \quad (9)$$

This is an optimization problem in two sets of variables with no mixed constraints. Furthermore, in our application, if one fixes either $a$ or $b$ and optimizes the other, then the resulting subproblem is much easier to solve than Problem (9) in its entirety. (Specifically, if one fixes $\Delta$ and minimizes $B$ in Problem (2), then one obtains Problem (4), for which Theorem 2 provides an explicit formula for a global solution. Alternatively, if one fixes $B$ and minimizes $\Delta$, then one obtains a convex problem.) Nonlinear programs with these characteristics are sometimes called problems whose variables separate, e.g., by Golub and Pereyra [8]. Parks [16], however, has argued persuasively for the superiority of the term reducible. Her comprehensive study of reducible nonlinear programming contains many valuable references. Following Trosset [28], we describe two important strategies for exploiting reducible structure. Various studies have suggested that both strategies inevitably improve on methods that fail to exploit the structure of reducible nonlinear programs.

#### 3.1 Variable Alternation

Variable alternation is the simple optimization strategy described in Figure 1. It has most commonly been used when each subproblem is projection into a subspace, in which case it is often called the Method of Alternating (orthogonal) Projections (MAP). MAP was first studied, in 1933, by von Neumann [30], who considered the problem of projecting into the intersection of two closed linear subspaces of a Hilbert space. In this setting, alternately projecting into each subspace converges to projection into their intersection. A recent survey of the MAP literature was made by Deutsch [6].

A more relevant use of variable alternation was made by Cheney and Goldstein [1], who considered the problem of minimizing the distance between two closed convex sets in Hilbert space, say $K_1$ and $K_2$. Let $P_i$ denote projection into $K_i$. Cheney and Goldstein established sufficient conditions for the sequence $(P_1P_2)^k x$ to converge to a point in $K_1$ nearest $K_2$. Specifically, convergence
1. Fix \( a_0 \). Set \( b_0 = \text{argmin } f(a_0, b) \) and \( k = 1 \).

2. Do until convergence:
   
   (a) \( a_k = \text{argmin } f(a, b_{k-1}) \)
   
   (b) \( b_k = \text{argmin } f(a_k, b) \)

   (c) \( k = k + 1 \)

Figure 1: A variable alternation strategy.

is assured either if one set is compact or if one set is finite-dimensional and the distance between the sets is attained.

Because of its simplicity and the fact that it produces a nonincreasing sequence of objective function values, variable alternation has appealed to a great many statisticians. A survey of variable alternation methods in statistics was made by de Leeuw [4]. Variable alternation has been widely used in algorithms for multidimensional scaling, most notably by Takane, Young and de Leeuw [22], in which context it is usually called the method of alternating least squares. The Data Box Algorithm of Glunt, Hayden and Raydan [7] applies variable alternation to a distance geometry problem with bound constraints.

Under very weak conditions, the convergence theory of Zangwill [32] can usually be exploited to establish that every accumulation point of a sequence produced by variable alternation is a fixed point, i.e. a pair \((a_\ast, b_\ast)\) for which \( a_\ast = \text{argmin } f(a, b_\ast) \) and \( b_\ast = \text{argmin } f(a_\ast, b) \). However, there is no general guarantee that \((a_\ast, b_\ast)\) solves Problem (9). For instance, Trosset [28] constructed a simple example in which a global maximizer was fixed under variable alternation.

In practice, variable alternation often does converge to a local minimizer. Unfortunately, the convergence rate of these methods is typically linear and often painfully slow.

### 3.2 Variable Reduction

The optimization strategy described in Figure 2 is usually employed when one of the subproblems can be solved explicitly. By a variable reduction method, we mean any method for solving

\[
\begin{align*}
\text{minimize} & \quad \bar{f}(a) \\
\text{subject to} & \quad a \in A.
\end{align*}
\]

In contrast to variable alternation, variable reduction possesses the following property:

**Theorem 3** If \( a_\ast \) is a global (local) minimizer of Problem (10), then \((a_\ast, b(a_\ast))\) is a global (local) minimizer of Problem (9).

**Proof:** If \( a_\ast \) is a global minimizer of Problem (10), then

\[
f(a_\ast, b(a_\ast)) = \bar{f}(a_\ast) \leq \bar{f}(a) = f(a, b(a)) \leq f(a, b)
\]

for all \((a, b) \in A \times B\). If \( a_\ast \) is a local minimizer of Problem (10), then there exists a neighborhood \( N(a_\ast) \) in which \( \bar{f}(a_\ast) \leq \bar{f}(a) \) and (11) holds for all \((a, b) \in N(a_\ast) \times B\). \( \square \)
1. For a fixed, define the value function by \( b(\bar{a}) = \arg\min f(\bar{a}, b) \).

2. Minimize the variable projection functional \( \bar{f}(a) = f(a, b(a)) \).

Figure 2: A variable reduction strategy.

Variable reduction eliminates one set of variables, albeit at the cost of complicating the objective function. A simple example examined by Trosset [28] illustrates that variable reduction can be considerably more efficient than variable alternation. The potential difficulty with variable reduction is that the value function may not be differentiable, although when it is the derivatives of \( \bar{f} \) usually have a very simple relation to the derivatives of \( f \). Generalized differentiability of value functions has been the subject of extensive investigation in nonsmooth optimization, e.g., Section 6.5 in [2].

For the special case of the semilinear least squares problem, differentiating the value function entails differentiating the Moore-Penrose pseudoinverse of a matrix. This was accomplished by Golub and Pereyra [8], who compared the performance of Gauss-Newton algorithms on the full and reduced forms of this problem. Ruhe and Wedin [18] determined that variable alternation on this problem exhibits linear convergence, whereas Gauss-Newton on the reduced problem exhibits superlinear convergence if Gauss-Newton on the full problem does.

Variable reduction has also been used in algorithms for multidimensional scaling, most notably by Kruskal [11, 12]. Variable reduction was first applied to the so-called additive constant problem by Saito [19]; de Leeuw and Heiser [5] and Trosset, Baggerly and Pearl [29] subsequently suggested using the variable projection functional \( F_p \circ \tau \). Trosset [27] proposed a formulation of nonmetric multidimensional scaling using \( F_p \circ \tau \) and Trosset [26, 28, 25] suggested variable reduction for distance geometry problems with bound constraints.

4 Optimization by Variable Alternation

Henceforth, we restrict attention to \( B_n = \Omega_n(K) \) for \( K \) a closed convex subset of \( (3) \), i.e., we study

\[
\begin{align*}
\text{minimize} & \quad f(\Delta, B) = \| \tau(\Delta) - B \|_F^2, \\
\text{subject to} & \quad \Delta \in C_n, \\
& \quad B \in \Omega_n(K).
\end{align*}
\]

(12)

In this section we establish some consequences of applying variable alternation to Problem (12). Our analysis requires an additional hypothesis:

**Assumption 1** For any \( \Delta^0 \in C_n \), the level set \( L_K(\Delta^0) = \{ \Delta \in C_n : F_K \circ \tau(\Delta) \leq F_K \circ \tau(\Delta^0) \} \) is bounded.

In the applications to which we have alluded, Assumption 1 can be established by demonstrating that \( \lim_{k \to \infty} F_K \circ \tau(\Delta^k) = \infty \) if \( \{\Delta^k\} \) is unbounded.
The ensuing discussion requires a precise definition of variable alternation on Problem (12). The variable alternation subproblems are Problem (4) for $\Delta \in \mathcal{C}_n$ fixed and

\begin{align}
\text{minimize} & \quad \| \tau(\Delta) - B \|^2_F \\
\text{subject to} & \quad \Delta \in \mathcal{C}_n
\end{align}

for $B \in \Omega_n(K)$ fixed. Notice that Problem (13) is the strictly convex problem of projecting $B$ into the closed convex set $\tau(\mathcal{C}_n)$.

Let $M_1$ be any function that assigns a single global solution of Problem (4) to each fixed $\Delta \in \mathcal{C}_n$. Let $M_2$ be any function that assigns the unique solution of Problem (13) to each fixed $B \in \Omega_n(K)$.

**Definition 6** By variable alternation on Problem (12) from $\Delta^0 \in \mathcal{C}_n$, we mean the method of generating from $\Delta^0$ the sequence $\{(\Delta^k, B^k)\}$, where $B^k = M_1(\Delta^k)$ and $\Delta^{k+1} = M_2(B^k)$. Furthermore, we say that $(\Delta^*, B^*)$ is a fixed point of variable alternation on Problem (12) if $B^* = M_1(\Delta^*)$ and $\Delta^* = M_2(B^*)$.

We begin by considering the case of $\Omega_n(K) = \Omega_n$, in which case variable alternation on Problem (12) consists of alternating projections between the closed and convex sets $\tau(\mathcal{C}_n)$ and $\Omega_n$.

**Theorem 4** If $\Omega_n(K) = \Omega_n$, then variable alternation on Problem (12) from any $\Delta \in \mathcal{C}_n$ converges to a global minimizer of Problem (12).

**Proof:** Let $\{\Delta^k\}$ be the sequence of dissimilarity matrices constructed by variable alternation on Problem (12) from $\Delta^0$. Let

\begin{equation}
\bar{f}_1(\Delta) = f(\Delta, M_1(\Delta)) = \| \tau(\Delta) - M_1(\Delta) \|^2_F = F_K \circ \tau(\Delta).
\end{equation}

Because

\begin{equation}
\bar{f}_1(\Delta^{k+1}) = f(\Delta^{k+1}, B^{k+1}) \leq f(\Delta^{k+1}, B^k) \leq f(\Delta^k, B^k) = \bar{f}_1(\Delta^k),
\end{equation}

\{\Delta^k\} \subseteq \mathcal{L}_K(\Delta^0)$ and variable alternation on Problem (12) from $\Delta^0$ is equivalent to variable alternation from $\Delta^0$ on the following problem:

\begin{align}
\text{minimize} & \quad f(\Delta, B) \\
\text{subject to} & \quad \Delta \in \mathcal{L}_K(\Delta^0), \\
& \quad B \in \Omega_n(K).
\end{align}

Furthermore, it is transparent that the global minimizers of Problems (16) and (12) are identical.

By Assumption 1, the closed set $\mathcal{L}_K(\Delta^0)$ is bounded, hence compact. Therefore, it follows from Theorem 4 in [1] that variable alternation on Problem (16) from $\Delta^0$ converges to a global minimizer of Problem (16). $\square$

We now consider the case of general $\Omega_n(K)$. Because Problem (12) is not convex except when $\Omega_n(K) = \Omega_n$, we obtain a weaker result.

**Theorem 5** Any sequence of dissimilarity matrices $\{\Delta^k\}$ obtained by applying variable alternation to Problem (12) will have at least one accumulation point. If $\Delta^*$ is an accumulation point of $\{\Delta^k\}$ and $B^* = M_1(\Delta^*)$, then $(\Delta^*, B^*)$ is a fixed point of variable alternation on Problem (12).

**Proof:** Because Problem (4) may have multiple solutions, it is convenient to apply the theory of point-to-set maps. Insofar as we can do so consistently, we adopt the terminology and notation of Hogan [10] and Zangwill [32].
Following Hogan [10], let \( \Omega \) denote the constant point-to-set map defined by \( \Omega(\Delta) \equiv \Omega_n(K) \). Since \( \Omega_n(K) \) is a closed set, it is easily verified that \( \Omega \) is continuous as a point-to-set map. Next let

\[
\nu(\Delta) = \inf \{ f(\Delta, B) : B \in \Omega_n(K) \} = \bar{f}_1(\Delta) = F_K \circ \tau(\Delta)
\]

and

\[
\bar{M}_1(\Delta) = \{ B \in \Omega(\Delta) : f(\Delta, B) \leq \nu(\Delta) \}.
\]

Since \( \Omega \) is a continuous map and \( f \) is a continuous function, it follows from Theorem 8 in [10] that the point-to-set map \( \bar{M}_1 \) is closed. (An elementary proof of this fact is possible, but the notation is cumbersome.)

Analogously, let \( \Theta \) denote the constant point-to-set map defined by \( \Theta(B) = C_n \). Like \( \Omega \), \( \Theta \) is continuous. Let

\[
w(B) = \inf \{ f(\Delta, B) : \Delta \in \Theta(B) \}
\]

and

\[
M_2(B) = \{ \Delta \in \Theta(B) : f(\Delta, B) \leq w(B) \}.
\]

Like \( \bar{M}_1 \), \( M_2 \) is a closed map. Moreover, we have already remarked that \( M_2(B) \) contains a single element, so that \( M_2 \) is in fact a function.

Following Zangwill [32], we now define an algorithmic point-to-set map \( A \) by \( A = M_2 \circ \bar{M}_1 \). By construction, \( \Delta^{k+1} \in A(\Delta^k) \). Suppose that \( \Delta^+ \in A(\Delta^k) \), in which case \( \Delta^+ = M_2(B^c) \) for some \( B^c \in \bar{M}_1(\Delta^c) \). Then, as in (15),

\[
\nu(\Delta^c) = f(\Delta^c, B^c) \geq f(M_2(B^c), B^c) = f(\Delta^+, B^c) \geq \nu(\Delta^+).
\]

Because \( M_2(B^c) \) is unique, the first inequality in (17) is strict if and only if \( \Delta^+ = M_2(B^c) = \Delta^c \).

Thus, either \( A(\Delta^c) = \{ \Delta^c \} \) or \( \Delta^c \not\in A(\Delta^c) \). We define \( \Delta^c \) to be a solution of \( A \) if \( A(\Delta^c) = \{ \Delta^c \} \) and note that \( \nu(\Delta^c) > \nu(\Delta^+) \) if \( \Delta^c \) is not a solution.

By Assumption 1, the closed set \( \mathcal{L}_K(\Delta^0) = \{ \Delta \in C_n : \nu(\Delta) \leq \nu(\Delta^0) \} \) is bounded, hence compact. Since it follows from (17) that \( \nu(\Delta^k) \leq \nu(\Delta^0) \) if \( \Delta^k \) is generated by \( A \), we conclude that any sequence generated by \( A \) must lie in a compact set and hence must have an accumulation point.

Since \( \nu(\Delta) \) is the distance between \( \Delta \) and \( \Omega_n(K) \), the closed set \( \bar{M}_1(\Delta) \) must be bounded, hence compact. It follows that, if \( \Delta^j \to \Delta \), then \( \cap_j \bar{M}_1(\Delta^j) \) is contained in a compact set. Hence, if \( B^j \in \bar{M}_1(\Delta^j) \), the sequence \( \{ B^j \} \) must contain a convergent subsequence. We can therefore apply Lemma 4.2 in [32] to conclude that \( A = M_2 \circ \bar{M}_1 \), the composition of closed maps, must itself be closed.

We now apply Convergence Theorem A in [32] to conclude that any accumulation point \( \Delta^* \) of any sequence generated by \( A \) must satisfy \( A(\Delta^*) = \{ \Delta^* \} \). For any choice of \( M_1 \), we then have \( \Delta^* = M_2 \circ M_1(\Delta^*) \). So \( (\Delta^*, M_1(\Delta^*)) \) is a fixed point of variable alternation on Problem (12).

In practice, variable alternation on Problem (12) does seem to find solutions. However, it converges very slowly.

## 5 Optimization by Variable Reduction

In principle, there are two distinct variable reduction strategies for solving Problem (12). First, consider the variable projection functional

\[
\bar{f}_2(B) = f(M_2(B), B) = \| \tau(M_2(B)) - B \|_p^2.
\]
Because $\tau(M_2(B))$ is the projection of $B$ into the closed convex set $C_n$, $\tilde{f}_2$ is a convex function. An explicit formula for $M_2(B)$ will rarely be available; however, in most applications, Problem (13) will be a quadratic programming problem and therefore $M_2(B)$ can be computed reasonably efficiently. This approach results in the following semidefinite programming problem:

$$\begin{align*}
\text{minimize} & \quad \tilde{f}_2(B) \\
\text{subject to} & \quad B \in \Omega_n(K).
\end{align*}$$

When $\Omega_n(K) = \Omega_n$, Problem (18) is convex. We are concerned with situations in which $\Omega_n(K)$ is a nonconvex subset of $\Omega_n$, particularly $\Omega_n(K) = \Omega_n(p)$. Tarazaga and Trosset [23] studied methods of managing the constraint $B \in \Omega_n(p)$. They reparametrized $B$ by writing $B = XX'$ for $X \in \mathbb{R}^{n \times p}$, then introduced a penalty function to remove the resulting indeterminancy in the representation of $B$. Unfortunately, this approach destroys many of the pleasant properties of Problem (18). Thus, the development of semidefinite programming methods that could manage rank restrictions directly would be of enormous value. In the absence of such methods, we explore an alternative approach.

Consider the variable projection functional defined by (14). Henceforth, we restrict attention to the case $\Omega_n(K) = \Omega_n(p)$ and consider the problem

$$\begin{align*}
\text{minimize} & \quad F_p \circ \tau(\Delta) \\
\text{subject to} & \quad \Delta \in \mathcal{C}_n.
\end{align*}$$

This problem has a closed convex feasible set, typically defined by linear constraints. (For example, in applications to computational chemistry, one usually imposes simple bound constraints on the $\delta_{ij}$.) Here, in contrast to Problem (18), our concern lies with the objective function. Fortunately, it behaves nicely:

**Theorem 6** Let $\Delta$ be a dissimilarity matrix and let $T = \tau(\Delta)$ have eigenvalues $\lambda_1(T) \geq \cdots \geq \lambda_n(T)$. Then $F_p \circ \tau$ is continuously differentiable at $\Delta$, unless $\lambda_p(T) = \lambda_{p+1}(T) \geq 0$.

**Proof:** Because $\tau$ is linear, it suffices to show that $F_p$ is continuously differentiable at $T$. From (8), we see that $F_p$ depends on $T$ only through its eigenvalues.

Let $Z$ denote an arbitrary $n \times n$ symmetric matrix and consider the analytic arc $T + \epsilon Z$ through $T$. By a famous theorem of Rellich [17], the eigenvalues of $T + \epsilon Z$ can be parametrized as analytic functions of $\epsilon$ in a neighborhood of $\epsilon = 0$. This parametrization may not preserve the ordering of the eigenvalues throughout the neighborhood; however, if $\lambda_p(T) > \lambda_{p+1}(T)$, then the neighborhood can be chosen small enough that the sets of the $p$ largest and the $n - p$ smallest eigenvalues are preserved within it. Then

$$F_p(T + \epsilon Z) = \sum_{i=1}^{p} \phi(\lambda_i(T + \epsilon Z)) + \sum_{i=p+1}^{n} \psi(\lambda_i(T + \epsilon Z))$$

is continuously differentiable in a neighborhood of $\epsilon = 0$, i.e., the directional derivative

$$F_p'(T)(Z) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} [F_p(T + \epsilon Z) - F_p(T)]$$

exists and is continuous.
If \( \lambda_p(T) = \lambda_{p+1}(T) < 0 \), then let \( q \) be such that \( \lambda_q(T) \) is the smallest strictly positive eigenvalue of \( T \). Since \( \phi(t) = \psi(t) \) when \( t < 0 \), there exists a neighborhood of \( \epsilon = 0 \) in which we can rewrite (20) as

\[
F_p(T + \epsilon Z) = \sum_{i=1}^q \phi(\lambda_i(T + \epsilon Z)) + \sum_{i=q+1}^n \psi(\lambda_i(T + \epsilon Z))
\]

with \( \lambda_q(T) > \lambda_{q+1}(T) \). Then the same argument again establishes the existence and continuity of the directional derivative. \( \Box \)

Formally, it is not difficult to compute the gradient of \( F_p \) at \( T \). We parametrize \( B \in \Omega_n(p) \) by \( B = XX' \), where \( X \in \mathbb{R}^{n \times p} \), and let \( f(T, X) = \|T - XX'\|^2 \). Let \( B^*(T) \in \Omega_n(p) \) denote a solution of Problem (4) with \( \Omega_n(K) = \Omega_n(p) \) and write \( B^*(T) = X(T)X(T)' \). Letting \( g_1(T) = T \), \( g_2(T) = X(T) \), and \( g = (g_1, g_2) \), we obtain \( F_p(T) = f \circ g(T) \), to which we will apply the chain rule. Notice that \( \nabla g_1(T) = I_n \) and that \( X(T) \) is a minimizer of the unconstrained optimization problem

\[
\text{minimize} \quad f(T, X) \\
\text{subject to} \quad X \in \mathbb{R}^{n \times p},
\]

so that \( \nabla_X f(g_1(T), g_2(T)) = 0 \). Then, if \( \nabla g_2(T) \) exists,

\[
[\nabla F_p(T)]' = [\nabla f \circ g(T)]' \\
= [\nabla f (g_1(T), g_2(T))][\nabla g_1(T)]' + [\nabla_X f(g_1(T), g_2(T))][\nabla g_2(T)]'
\]

\[
= 2[T - g_2(T)g_2(T)]' \\
= 2[T - B^*(T)],
\]

which is a special case of Corollary 4.5 in [16].

It is instructive to consider why \( F_p \) is not everywhere differentiable. If \( T \) is such that Problem (4) with \( \Omega_n(K) = \Omega_n(p) \) has a unique solution, \( B^*(T) \), then \( F_p \) is differentiable at \( T \) and \( -\nabla F_p(T) \) is the direction of steepest descent toward \( B^*(T) \). If \( T \) is such that Problem (4) with \( \Omega_n(K) = \Omega_n(p) \) has multiple solutions, then \( F_p \) is not differentiable at \( T \). In this case, however, one can still compute the last expression in (21) for each solution, and each \(-2[T - B^*(T)]\) so obtained is the direction of steepest descent toward the corresponding solution. Thus, (21) provides meaningful information even when it is not formally the gradient of \( F_p \).

To facilitate developing algorithms for solving Problem (19), we now reinterpret \( F_p \circ \tau \) as a function of the \( m = n(n-1)/2 \) subdiagonal elements of the dissimilarity matrix \( \Delta \) and \( C_n \) as a closed convex subset of \( \mathbb{R}^m \). Then, writing

\[
\delta = (\delta_{i_1j_1}, \ldots, \delta_{i_mj_m})' \in \mathbb{R}^m,
\]

we can rewrite Problem (19) as:

\[
\text{minimize} \quad F_p \circ \tau(\delta) \\
\text{subject to} \quad \delta \in C_n.
\]

(22)

When it exists, it is not difficult to compute the gradient of \( F_p \circ \tau \) with respect to the \( \delta_{i_1j_1} \). We define \( n \times n \) matrices \( E^{rs} = [e_{ij}^{rs}] \) by setting \( e_{ii}^{rs} = e_{jj}^{rs} = 1 \) and all other elements equal to zero. From (1), it is easily seen that

\[
\frac{\partial}{\partial \delta_{i_1j_1}} \tau(\delta) = \tau(E^{i_1j_1}).
\]
Applying the chain rule yields
\[
\frac{\partial}{\partial \delta_{i,k,j,k}} F_p \circ \tau(\delta) = \left( \nabla F_p(\tau(\delta)), \tau(E_{i,k,j,k}) \right)_F.
\] (23)

Then, using (21) with \( S = B^*(\tau(\delta)) - \tau(\delta) \) and summing over indices denoted by \(+\), some computation simplifies (23) to
\[
\frac{\partial}{\partial \delta_{i,k,j,k}} F_p \circ \tau(\delta) = \frac{2}{n^2} S_{++} - \frac{2}{n} (S_{i+k} + S_{j+k}) + 2 S_{i,j,k,k}.
\] (24)

Because the feasible set for Problem (22) is closed and convex, it is natural to contemplate a gradient projection method for its solution. Let \( \delta^0 \in \mathcal{C}_n \) be given and let \( P \) denote projection into \( \mathcal{C}_n \). Referring to (21), we define a gradient projection method for Problem (22) to mean the construction of a sequence
\[
\delta^{k+1} = P\left( \delta^k - 2t_k \left[ \tau(\delta^k) - B^*(\tau(\delta^k)) \right] \right),
\] (25)

where \( t_k \) is obtained by a line search, i.e., by solving the univariate optimization problem
\[
\begin{align*}
\text{minimize} & \quad F_p \circ \tau \left[ P\left( \delta^k - 2t \left[ \tau(\delta^k) - B^*(\tau(\delta^k)) \right] \right) \right] \\
\text{subject to} & \quad t \geq 0.
\end{align*}
\] (26)

As we have noted, there is some ambiguity in the definition of \( B^*(\tau(\delta)) \) when Problem (4) has multiple solutions at \( T = \tau(\delta) \). In practice, one might perform a single line search using an arbitrary solution or perform separate line searches using each solution and take the best \( \delta \) so obtained.

We now borrow from McCormick and Tapia [15]:

**Definition 7** Given \( \delta \in \mathcal{C}_n \), let \( \overline{\mathcal{A}}(\delta) \) denote the closure of the tangent cone to \( \mathcal{C}_n \) at \( \delta \) and let \( P_\delta \) denote projection into \( \overline{\mathcal{A}}(\delta) \). Then \( \delta \) is a constrained stationary point of Problem (22) if
\[
P_\delta [-\nabla F_p \circ \tau(\delta)] = 0.
\]

**Theorem 7** Suppose that \( \delta^* \) is an accumulation point of the gradient projection sequence defined by (25). If \( F_p \circ \tau \) is differentiable at \( \delta^* \), then \( \delta^* \) is a constrained stationary point of Problem (22). If Assumption 1 is satisfied, then the sequence will have at least one accumulation point.

**Proof:** Because \( \{ F_p \circ \tau(\delta^k) \} \) is a nonincreasing sequence, Assumption 1 implies that the sequence \( \{ \delta^k \} \) is contained in a compact set and therefore has an accumulation point.

If \( F_p \circ \tau \) is differentiable at the accumulation point \( \delta^* \), then the argument used to establish Theorem 1 in [15] is valid in a neighborhood of \( \delta^* \) and establishes that \( \delta^* \) is a constrained stationary point of Problem (22). \( \square \)

Theorem 7 is stronger than Theorem 5 in the following sense: if \( \delta^* \) is a constrained stationary point of Problem (22), then \( (\tau(\delta^*), B^*(\tau(\delta^*))) \) is a fixed point under variable alternation on the corresponding Problem (12), but not conversely. Furthermore, although gradient projection methods are linearly convergent, sequences constructed by gradient projection on Problem (22) tend to converge much more rapidly than sequences constructed by gradient projection on Problem (12).
6 Discussion

We have presented several formulations of the general problem of minimizing the (squared) distance (in Frobenius norm) between certain closed convex sets of symmetric \( n \times n \) matrices and certain subsets of the symmetric positive semidefinite \( n \times n \) matrices. The former sets are obtained by linear transformation of closed convex sets of dissimilarity matrices. In most applications, the latter sets are defined by rank restrictions. At present, our preferred formulation of such problems is Problem (22).

Let us reflect on some of the properties possessed by Problem (22):

1. The number of variables, \( m = n(n-1)/2 = O(n^2) \), is potentially huge. In applications to computational chemistry, \( n \) may be several hundreds, if not thousands.

2. The objective function is (usually) continuously differentiable and analytic derivatives are available. The cost of a function or gradient evaluation is effectively the cost of computing the spectral decomposition of a symmetric \( n \times n \) matrix.

3. The objective function does not have a great deal of curvature near solutions. This may be inferred from the well-known fact that classical multidimensional scaling is extremely stable under perturbations of the dissimilarity data, a phenomenon that has been studied by Mardia [13] and by Sibson [21].

4. Analytic Hessians are not available. Typically, the Hessian matrix is completely dense and \( m = m(m+1)/2 = O(n^4) \) variables are required to specify it (or an approximation thereof). Finite-difference approximations of the Hessian matrix are extremely expensive to compute.

5. The feasible set is closed and convex. In most applications, it is polyhedral. In applications to computational chemistry, it is defined by specifying simple bounds on each variable.

There is an evident tension between certain of these properties, in that we would like to exploit information about the curvature of the objective function to construct algorithms with fast local convergence, but such information is expensive to obtain and to manage. Because of this tension, the development of efficient algorithms for solving Problem (22) is likely to be specific to the application and is beyond the scope of this report. We have suggested a gradient projection method for general use. We are currently studying interior-point methods for the bound-constrained problems that arise in computational chemistry.

This report has emphasized local theory, which is usually the most for which one can hope. It is worth noting, however, that the formulations that we have presented appear to be extremely well-behaved. In a study of distance matrix completion problems, Trosset [26] reported that local searches started from random feasible points always seemed to converge to global solutions. In further numerical experiments, to be described in future reports on specific applications, we have thus far encountered very few difficulties with nonglobal solutions. These experiences commend our formulations of these problems and offer hope that it may be possible to develop global optimization strategies for solving them. Such developments would have profound implications for the disciplines from which these problems originate.
Acknowledgments

This research was supported by grant DMS-9622749 from the National Science Foundation and by the W. M. Keck Center for Computational Biology under Medical Informatics Training grant 1T1507093 from the National Library of Medicine. I have benefited greatly from discussions with a number of colleagues, particularly Jan de Leeuw, Timothy Havel, Yoshio Takane, Richard Tapia and Pablo Tarazaga.

References


