Better initial configurations for metric multidimensional scaling

Samuel W. Malone\textsuperscript{a,1}, Pablo Tarazaga\textsuperscript{b,2}, Michael W. Trosset\textsuperscript{c,*3}

\textsuperscript{a}Department of Mathematics, Duke University, Durham, NC 27708-0586, USA
\textsuperscript{b}Department of Mathematics, University of Puerto Rico, Mayagüez, PR 00681-5000, USA
\textsuperscript{c}Department of Mathematics, College of William & Mary, P.O. Box 8795, Williamsburg, VA 23185-8795, USA

Received 1 February 2002; received in revised form 1 March 2002

Abstract

Multidimensional scaling (MDS) is a collection of data analytic techniques for constructing configurations of points from dissimilarity information about interpoint distances. Two popular measures of how well the constructed distances fit the observed dissimilarities are the raw stress and sstressed criteria, each of which must be minimized by numerical optimization. Because iterative procedures for numerical optimization typically find local minimizers that may not be global minimizers, the choice of an initial configuration from which to begin searching for an optimal configuration is crucial. A popular choice of initial configuration is the classical solution of Torgerson (Psychometrika 17 (1952) 401). Results from the theory of distance matrices are exploited to derive two alternatives, each guaranteed to be at least as good as the classical solution, and empirical evidence is presented that they are usually substantially better.

\(\text{c} \, \text{©} \, 2002 \text{ Elsevier Science B.V. All rights reserved.}

\textit{MSC:} 62H25; 51K99; 15A18

\textit{Keywords:} Distance matrices; Distance geometry; Spectral decomposition; Low rank approximations; Stress and sstress criteria; Global optimization

\(\text{c} \, \text{©} \, 2002 \text{ Elsevier Science B.V. All rights reserved.}

\textit{E-mail addresses:} swm3@duke.edu (S.W. Malone), tarazaga@cs.uprm.edu (Pablo Tarazaga), trosset@math.wm.edu (Michael W. Trosset).

1 Supported by Grant DMS-9619577 from the National Science Foundation.
2 Supported by Grants HRD-9450448 and CDA-94117362 from the National Science Foundation.
3 Supported by Grants DMS-9622749 and DMS-9619577 from the National Science Foundation.
1. Introduction

Multidimensional scaling (MDS) is a general term for techniques that construct configurations of points in a target metric space (typically \( p \)-dimensional Euclidean space) from information about interpoint distances. Originally developed by psychometricians, MDS has become a standard technique in multivariate data analysis and is widely used in a variety of disciplines. Its popularity illustrates the importance of data visualization, a prominent theme in modern statistics.

In the case of two-way MDS, data are specified in the form of a dissimilarity matrix, i.e. a matrix \( A = [\delta_{ij}] \) such that \( \delta_{ij} \geq 0, \delta_{ii} = 0, \) and \( \delta_{ij} = \delta_{ji}. \) A dissimilarity matrix is often obtained by measuring human perceptions of difference, in which case the goal of MDS is the construction of a “psychological space” that represents stimuli not as they are but as they are perceived. Dissimilarity matrices studied in Section 5 include examples from psychophysics (how people perceive colors), marketing (how consumers perceive colas), and political science (how people perceive nations). In most applications the target dimension \( p \) is small in order to facilitate visualization.

A slightly different application of MDS (under the heading distance geometry) occurs in computational chemistry and structural molecular biology. In this application, upper and lower bounds on interatomic distances are derived from chemical structure (see Trosset and Phillips (1999) for a detailed description) and nuclear magnetic resonance (NMR) spectroscopy; the goal is to reconstruct three-dimensional molecular structure. Trosset (1998) has discussed adapting MDS to this end.

For a configuration of points \( x_1, \ldots, x_n \in \mathbb{R}^p, \) the \( n \times p \) configuration matrix \( X \) is the matrix whose rows are the \( x_i \). From \( X \) it is easy to compute the Euclidean interpoint distance matrix \( D(X) = (d_{ij}) \). The goal of two-way metric MDS is the construction of a configuration matrix for which the interpoint distances \( d_{ij} \) approximate the dissimilarities \( \delta_{ij}. \) Two popular ways of measuring the discrepancy between a distance matrix and a dissimilarity matrix are the raw stress and ssstress criteria. The former is based on the squared errors between the distances and dissimilarities; the latter is based on the squared errors between the squared distances and squared dissimilarities. Let

\[
\rho_r(D, \Delta) = \sum_{ij} w_{ij} [(d_{ij})^r - (\delta_{ij})^r]^2,
\]

where the \( w_{ij} \) are nonnegative weights. Given a fixed dissimilarity matrix \( \Delta, \) write \( D = D(X) \) and let \( \sigma_r(X) = \rho_r(D(X), \Delta); \) then \( \sigma_1 \) is the raw stress criterion and \( \sigma_2 \) is the slightly less popular raw ssstress criterion.

In practice, one often sets each \( w_{ij} = 1. \) This is the only case that we consider. However, one can use the weights either to accommodate missing data (by setting the appropriate \( w_{ij} = 0 \)) or to weight more reliably measured dissimilarities more heavily.

Having selected an error criterion \( \sigma_r, \) MDS attempts to find a configuration matrix \( X \) that minimizes \( \sigma_r. \) Minimizers must be computed by an iterative algorithm for numerical optimization. A survey of some of the more efficient algorithms for minimizing \( \sigma_1 \) and \( \sigma_2 \) was made by Kearsley et al. (1998). These algorithms find local minimizers. To improve the chance that the algorithm will converge to a global minimizer, the user is advised to choose a good initial configuration. The conventional choice—the
default initial configuration used in many implementations—is the configuration constructed by classical MDS. Classical MDS involves nothing more complicated than a single spectral decomposition of a symmetric matrix and usually provides an excellent configuration from which to begin minimizing $\sigma_r$.

In the 1990s, the MDS community became increasingly concerned with the possibility that algorithms for minimizing stress or stress can be trapped by nonglobal minimizers. Some researchers, e.g. Groenen (1993) and Groenen and Heiser (1996), argued that nonglobal minimizers of the raw stress criterion are common and recommended using global optimization algorithms. In contrast, Kearsley et al. (1998) demonstrated that the popular SMACOF-I algorithm (Heiser and de Leeuw, 1986) for minimizing stress tends to terminate prematurely, creating the false impression that a local minimizer has been found. They recommended using more sophisticated, second-order methods.

Because the only evidence for nonglobal minimizers of stress and stress had been empirical, the results in Kearsley et al. (1998) offered the tantalizing possibility that nonglobal minimizers do not exist after all. This possibility was eliminated by Trosset and Mathar (1997), who analytically verified the existence of a nonglobal minimizer of the raw stress criterion. The prevalence of such minimizers is a matter that requires careful empirical investigation.

Whatever algorithm one prefers for minimizing stress or stress, its performance can only benefit by starting from a configuration that is close to a global minimizer. The sections that follow describe simple ways to compute excellent initial configurations from which to begin minimizing. In Section 2 we briefly describe classical MDS. The remaining sections present research that has only been disseminated in two conference proceedings papers (Tarazaga and Trosset, 1998; Malone and Trosset, 2000a) and a technical report (Malone and Trosset, 2000b) In Section 3 we formulate an approach that is guaranteed to outperform the classical solution in the special case $r = 2$ (stress). In Section 4 we derive a simple device for improving any initial configuration with respect to any choice of $r$. In Section 5 we present numerical results that suggest that both techniques often perform substantially better than the classical solution.

2. Classical MDS

Given $p \leq n$, let $D_p(n)$ denote the set of $p$-dimensional distance matrices, i.e. $D = (d_{ij}) \in D_p(n)$ if and only if there exists a configuration $x_1, \ldots, x_n \in \mathbb{R}^p$ such that $d_{ij} = \|x_i - x_j\|$. Given $D = [d_{ij}] \in D_p(n)$, let $D_r$ denote the matrix with entries $(d_{ij})^r$. Likewise, given a configuration $x_1, \ldots, x_n \in \mathbb{R}^p$, let $D_r(X)$ denote the matrix with entries $\|x_i - x_j\|^r$.

We define two closed sets in the space of symmetric $n \times n$ matrices. First, let $A_p(n)$ denote the set of $p$-dimensional squared distance matrices, i.e. $A_p(n) = \{D_2 : D \in D_p(n)\}$. Second, let $\Omega_p(n)$ denote the closed set of symmetric $n \times n$ matrices that are positive semidefinite and have rank no greater than $p$. Schoenberg (1935) and Young and Householder (1938) independently discovered a linear relation between $A_p(n)$ and $\Omega_p(n)$. This relation can be stated in terms of the cosine law.
Given $n$, let $e=(1,\ldots,1)'\in\mathbb{R}^n$. Given an $n\times n$ matrix $B=(b_{ij})$, let $b=(b_{11},\ldots,b_{nn})'\in\mathbb{R}^n$ denote the diagonal elements of $B$. Define a linear transformation $\kappa : \Omega_n(n) \rightarrow A_n(n)$ by $\kappa(B) = be' + eb' - 2B$. Then the following result is well-known:

**Theorem 1** (Embedding Theorem). $D_2 \in A_n(p)$ if and only if there exists $B \in \Omega_n(p)$ such that $\kappa(B) = D_2$.

The linear transformation $\kappa$ does not have a unique inverse. In fact, the null space of $\kappa$ is the set of all matrices of the form $e'x + xe'$ for some $x \in \mathbb{R}^n$. Let $I$ denote the $n \times n$ identity matrix. Given any $s \in \mathbb{R}^n$ such that $s'e=1$, the linear transformation $\tau_s$ defined by

$$
\tau_s(D) = -\frac{1}{2} (I-es')D(I-es')
$$

is an inverse of $\kappa$. Johnson and Tarazaga (1995) demonstrated that the image of $\tau_s$ is a face of the polyhedral cone $\Omega_n(n)$. The inverse transformation $\tau_1$ obtained by setting $s=e/n$ was introduced by Torgerson (1952). See Critchley (1988) for a detailed study of the properties of $\kappa$ and $\tau_1$.

Implicit in Torgerson’s (1952) and Gower’s (1966) pioneering studies of metric MDS was an optimization problem. Given a dissimilarity matrix $A = [\delta_{ij}]$, let $A_r$ denote the matrix with entries $(\delta_{ij})'$. Let $\| \cdot \|_F$ denote the Frobenius norm, i.e. the norm induced by the inner product $\langle A,B \rangle_F = \text{trace}(A'B)$. Then classical MDS can be defined by the optimization problem

$$
\text{minimize} \quad \| B - \tau_1(A_2) \|_F^2
$$

subject to $B \in \Omega_n(p).$ \hspace{1cm} (1)

This formulation of classical MDS has been explicitly discussed by Mardia (1978), by de Leeuw and Heiser (1982), and by Trosset (1997, 2002). The objective function is sometimes called the strain criterion.

The following explicit solution to Problem (1) is well-known:

**Theorem 2** (Classical MDS). Given $A$, let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the eigenvalues of $B=\tau_1(A_2)$ and let $v_1,\ldots,v_n$ denote the corresponding eigenvectors. Given $p \leq n$, let $\lambda_i^+ = \max(\lambda_i,0)$ for $i = 1,\ldots,p$. Then

$$
\hat{B} = \sum_{i=1}^{p} \lambda_i^+ v_i v_i'
$$

is a global minimizer of Problem 1. Furthermore, if $\hat{X}$ is the $n \times p$ configuration matrix whose $ith$ column is $(\lambda_i^+)^{1/2}v_i$, then

$$
\kappa(\hat{B}) = D_2(\hat{X}).
$$
Because the classical solution, $\hat{X}$, can be computed explicitly, it is often used as the initial configuration from which optimization of $\sigma_r$ commences.

In anticipation of Section 3, we now offer a slightly different formulation of classical MDS. Given a set $w_1, \ldots, w_p \in \mathbb{R}^n$ of $p$ orthonormal $n$-vectors, we introduce the polyhedral cone

$$C(w_1, \ldots, w_p) = \left\{ B : B = \sum_{i=1}^{p} \mu_w w_i w_i' ; \mu_i \geq 0 \right\}.$$ 

It follows from the spectral theorem for symmetric matrices that:

**Lemma 1.** The set $\Omega_n(p)$ is the union, over all sets of $p$ orthonormal $n$-vectors, of the polyhedral cones $C(w_1, \ldots, w_p)$.

Given

$$B^0 = \tau_1(A_2) = \sum_{i=1}^{n} \lambda_i v_i v_i',$$

we now substitute the $p$-dimensional cone $C(v_1, \ldots, v_p)$ for the $p(p+1)/2$-dimensional set $\Omega_n(p)$ in Problem (1), obtaining

$$\begin{align*}
\text{minimize} & \quad \|B - B^0\|_F^2 \\
\text{subject to} & \quad B \in C(v_1, \ldots, v_p).
\end{align*}$$

Despite the reduction in dimensionality, it follows from Theorem 2 that:

**Theorem 3.** If $B^*$ is a global minimizer of Problem (3), then $B^*$ is a global minimizer of Problem (1).

3. The sstress criterion

In this section we specialize to $r = 2$ and consider the problem of minimizing the raw sstress criterion,

$$\begin{align*}
\text{minimize} & \quad \|D_2 - A_2\|_F^2 \\
\text{subject to} & \quad D_2 \in \Lambda_n(p).
\end{align*}$$

We will exploit the linear relation between $\Lambda_n(p)$ and $\Omega_n(p)$ to develop analogies between Problems (4) and (1). Specifically, we attempt to obtain a formulation of Problem (4) that is analogous to our formulation of classical MDS in Problem (3). Although not entirely successful, our attempt leads to a new approximate solution.
The image of the polyhedral cone $C(w_1, \ldots, w_p)$ under the linear transformation $\kappa$ is the polyhedral cone

$$C_\kappa(w_1, \ldots, w_p) = \left\{ D : D = \sum_{i=1}^{p} \mu_i \kappa(w_i w'_i), \mu_i \geq 0 \right\}.$$ 

The following result is analogous to Lemma 1.

**Lemma 2.** The set $\Lambda_n(p)$ is the union, over all sets of $p$ orthonormal $n$-vectors, of the polyhedral cones $C_\kappa(w_1, \ldots, w_p)$.

**Proof.** Given $D_2 \in \Lambda_n(p)$, there exists $B \in \Omega_n(p)$ such that $D_2 = \kappa(B)$. Because $B \in \Omega_n(p)$, there exist eigenvalues $\mu_1, \ldots, \mu_p \geq 0$ and corresponding eigenvectors $w_1, \ldots, w_p$ such that

$$B = \sum_{i=1}^{p} \mu_i w_i w'_i.$$ 

By the linearity of $\kappa$,

$$D_2 = \kappa(B) = \sum_{i=1}^{p} \mu_i \kappa(w_i w'_i) \in C_\kappa(w_1, \ldots, w_p).$$

Given (2), we now substitute the $p$-dimensional cone $C_\kappa(v_1, \ldots, v_p)$ for the $p(p+1)/2$-dimensional set $\Lambda_n(p)$ in Problem (4), obtaining

$$\begin{align*}
\text{minimize} & \quad \|D_2 - A_2\|_F^2 \\
\text{subject to} & \quad D_2 \in C_\kappa(v_1, \ldots, v_p). 
\end{align*}$$

(5)

Unfortunately, Problem (5) is in general not equivalent to Problem (4). However, the following result suggests that a solution of the former may be a good approximation to a solution of the latter.

**Theorem 4.** Given (2), let $\hat{B}$ denote a global solution of Problem (1) and let $D_2^*$ denote a global solution of Problem (4). Then

$$\|D_2^* - A_2\|_F^2 \leq \|\kappa(\hat{B}) - A_2\|_F^2.$$ 

(6)

**Proof.** We have $\|D_2^* - A_2\|_F^2 \leq \|D_2 - A_2\|_F^2$ for every $D_2 \in C_\kappa(v_1, \ldots, v_p)$. But $\hat{B} \in C(v_1, \ldots, v_p)$; hence, $\kappa(\hat{B}) \in C_\kappa(v_1, \ldots, v_p)$ and (6) follows. 

Theorem 4 states that, measured by the raw stress criterion, a solution to Problem (5) is necessarily better than the classical solution. In fact, the results presented in Section 5 suggest that, typically, it is much better. To see that global solutions of
Problem (5) can be computed easily, it helps to rewrite Problem (5) in the following form:

\[
\text{minimize } f(\mu) = \left\| \sum_{i=1}^{p} \mu_i \kappa(v_i \theta^T_i) - A_2 \right\|_F^2
\]

subject to \( \mu_1, \ldots, \mu_p \geq 0. \) (7)

We see that Problem (7) is a strictly convex quadratic program in just \( p \) variables. The only constraints are bound constraints, so a variety of algorithms (e.g. \texttt{nlinmb} in S-Plus) are available for its solution.

4. Optimal dilations

Glunt et al. (1991) established the remarkable result that, for any fixed dissimilarity matrix, all of the interpoint squared distance matrices generated by stationary configurations of the raw stress criterion with unit weights lie on the surface of a sphere. We have extended this result from the special case of \( r = 2 \) to the general case of \( r > 0. \)

Given \( r > 0 \) and a fixed dissimilarity matrix \( A, \) we are interested in stationary configurations of the error criterion

\[
\sigma_r(X) = \|D_r(X) - A_r\|_F^2,
\]

i.e. in configuration matrices \( \tilde{X} \) for which \( \nabla \sigma_r(\tilde{X}) = 0. \) The following result has interesting implications for global optimization of \( \sigma_r. \)

**Theorem 5** (Generalized Critical Point Theorem). If \( \nabla \sigma_r(\tilde{X}) = 0, \) then

(i) \( \langle D_r(\tilde{X}), A_r \rangle_F = \|D_r(\tilde{X})\|_F^2; \)
(ii) \( \|D_r(\tilde{X})\|_F^2 + \|D_r(\tilde{X}) - A_r\|_F^2 = \|A_r\|_F^2; \)
(iii) If \( \nabla \sigma_r(\tilde{Y}) = 0, \) then \( \sigma_r(\tilde{X}) < \sigma_r(\tilde{Y}) \) if and only if \( \|D_r(\tilde{Y})\|_F < \|D_r(\tilde{X})\|_F; \)
(iv) \( \|D_r(\tilde{X}) - A_r/2\|_F = \|A_r/2\|_F. \)

**Proof.** Let

\[
g_r(t) = \sigma_r(t\tilde{X}) = \|D_r(t\tilde{X}) - A_r\|_F^2 = \|tD_r(\tilde{X}) - A_r\|_F^2
\]

\[= t^{2r} \|D_r(\tilde{X})\|_F^2 - 2t^r \langle D_r(\tilde{X}), A_r \rangle_F + \|A_r\|_F^2.\]

Differentiating, we obtain

\[
g_r'(t) = 2rt^{2r-1}\|D_r(\tilde{X})\|_F^2 - 2rt^{r-1} \langle D_r(\tilde{X}), A_r \rangle_F.\]
Because $\nabla \sigma_r(\vec{X}) = 0$, we have $g'_r(1) = 0$ and therefore

$$\langle D_r(\vec{X}), A_r \rangle_F = \|D_r(\vec{X})\|_F^2,$$

which is (i).

Next, applying (i), we observe that

$$\|D_r(\vec{X}) - A_r\|_F^2 = \|D_r(\vec{X})\|_F^2 - 2\langle D_r(\vec{X}), A_r \rangle_F + \|A_r\|_F^2$$

$$= \|A_r\|_F^2 - \|D_r(\vec{X})\|_F^2.$$

Rearranging terms produces (ii).

Now suppose that $\nabla \sigma_r(\vec{X}) = \nabla \sigma_r(\vec{Y}) = 0$. Applying (ii), we obtain

$$\sigma_r(\vec{X}) - \sigma_r(\vec{Y}) = \|D_r(\vec{X}) - A_r\|_F^2 - \|D_r(\vec{Y}) - A_r\|_F^2$$

$$= [\|A_r\|_F^2 - \|D_r(\vec{X})\|_F^2] - [\|A_r\|_F^2 - \|D_r(\vec{Y})\|_F^2]$$

$$= \|D_r(\vec{Y})\|_F^2 - \|D_r(\vec{X})\|_F^2,$$

from which (iii) follows immediately.

Finally,

$$\|D_r(\vec{X}) - A_r/2\|_F^2 = \|D_r(\vec{X})\|_F^2 - \langle D_r(\vec{X}), A_r \rangle_F + \|A_r/2\|_F^2$$

$$= \|A_r/2\|_F^2,$$

which is (iv).

Part (iv) of Theorem 5 states that, for any fixed dissimilarity matrix $A$, the inter-point distance matrices of all stationary configurations lie on the same sphere of radius $\|A/2\|_F$, centered at $A/2$. Theorem 5 does not address the existence of non-global minimizers, but part (iii) states that global minimizers are those stationary configurations of maximal norm.

We believe that what follows is the first successful attempt to exploit the geometry of the critical points of $\sigma_r$ for the purpose of minimizing $\sigma$. Suppose that a configuration matrix $X$ has been proposed as a possible minimizer of $\sigma$. The following result states that, by dilating $X$ so that (i) is satisfied, we necessarily decrease $\sigma$.

**Theorem 6 (Dilation Theorem).** Let $A$ be a fixed dissimilarity matrix, let $X$ be a fixed configuration matrix for which $\|D(X)\|_F > 0$, and consider the function $g_r : \mathbb{R} \to \mathbb{R}$ defined by $g_r(t) = \sigma_r(tX)$. Let

$$t^*_r = \left( \frac{\langle D_r(X), A_r \rangle_F}{\|D_r(X)\|_F^2} \right)^{1/r}.$$
Then

(i) \( D_r(t^*X) \) lies on the sphere described by (iv) in Theorem 5; and
(ii) \( t^* \) is a global minimizer of \( g_r \).

**Proof.** To establish (i), we compute

\[
\|D_r(t^*X) - A_r/2\|_F^2 = \|D_r(t^*X)\|_F^2 - \langle D_r(t^*X), A_r \rangle_F + \|A_r/2\|_F^2
\]

\[
= (t^*)^2 \|D_r(X)\|_F^2 - (t^*)^2 \langle D_r(X), A_r \rangle_F + \|A_r/2\|_F^2
\]

\[
= \frac{\langle D_r(X), A_r \rangle_F^2}{\|D_r(X)\|_F^2} - \frac{\langle D_r(X), A_r \rangle_F^2}{\|D_r(X)\|_F^2} + \|A_r/2\|_F^2
\]

\[
= \|A_r/2\|_F^2,
\]

which is (iv) in Theorem 5.

To minimize

\[
g_r(t) = \sigma_r(tX) = \|D_r(tX) - A_r\|_F^2 = \|t'D_r(X) - A_r\|_F^2
\]

\[
= t'^2 \|D_r(X)\|_F^2 - 2t'^2 \langle D_r(X), A_r \rangle_F + \|A_r/2\|_F^2,
\]

we first note that \( g_r(t) \to \infty \) as \( t \to \pm \infty \). Hence, it suffices to consider the stationary points of \( g_r \), i.e. the values of \( t \) at which

\[
g'_r(t) = 2rt^{r-1} \|D_r(X)\|_F^2 - 2r^{r-1} \langle D_r(X), A_r \rangle_F
\]

vanishes. By inspection, \( g'_r(t) = 0 \) if and only if either \( t = 0 \) or

\[
t' = \frac{\langle D_r(x), A_r \rangle_F}{\|D_r(X)\|_F^2} = (t^*)^r,
\]

and

\[
g_r(t^*) = \frac{\langle D_r(X), A_r \rangle_F^2}{\|D_r(X)\|_F^2} - 2 \frac{\langle D_r(X), A_r \rangle_F^2}{\|D_r(X)\|_F^2} + \|A_r\|_F^2 < \|A_r\|_F^2 = g_r(0).
\]

Noting that \( g_r(t) \) depends on \( t \) only through \( t' \), we conclude that \( t^* \) is a global minimizer of \( g_r \).  \( \square \)

Again we specialize to the case \( r = 2 \) and revisit the methods described in Section 5. Problem (7) was derived by replacing each \( \beta_i^+ \) in Theorem 2 with a freely varying \( \mu_i \). Suppose, instead, that we write \( \mu_i = t\beta_i^+ \) and allow \( t \in \mathbb{R} \) to vary. This results in
the objective function
\[ f(t\hat{\kappa}^+) = \left\| \sum_{i=1}^{p} t\hat{\lambda}_i \kappa(v_i, v'_i) - A_2 \right\|^2 = \left\| \kappa(t\hat{B}) - A_2 \right\|^2 = \left\| D_2(t\hat{X}) - A_2 \right\|^2 = g_2(t) \]
and the optimization problem
\[
\text{minimize} \quad g_2(t) \\
\text{subject to} \quad t \geq 0. \tag{8}
\]

Recognizing that Problem (8) is the problem of optimally dilating the classical solution, we can apply Theorem 6 to compute its solution explicitly, without recourse to numerical optimization. Because Problem (8) is a special case of Problem (7) with fewer degrees of freedom, we conclude that optimally dilating the classical solution for \( r = 2 \) is necessarily inferior to solving Problem (7). However, it is much easier to compute \( t^*_r \) in Theorem 6 than it is to solve Problem (7). Furthermore, the numerical results reported in Section 5 suggest that the optimal dilation of the classical solution is often competitive with the solution of Problem (7).

Finally, suppose that \( \Delta = [\delta_{ij}] \) is actually a distance matrix, i.e. that \( \Delta \in \mathbf{D}_n(n-1) \). Let \( \hat{X} \) denote the configuration matrix defined in Theorem 2 and write
\[ D(\hat{X}) = \hat{D} = [\hat{d}_{ij}] \]
Then it is well-known—see Meulman (1992) for discussion—that \( \hat{d}_{ij}^2 \leq \delta_{ij}^2 \). Because
\[ g_r(t) = \left\| t^r \hat{D}_r - A_r \right\|_F^2 = \sum_{ij} \left[ t^r (\hat{d}_{ij}^r) - (\delta_{ij})^r \right]^2, \]
we have the following result:

**Theorem 7.** Let \( \Delta \) be a fixed dissimilarity matrix. Given \( p \), let \( \hat{X} \) denote the configuration matrix defined in Theorem 2. Given \( r \), let \( t^*_r \) denote the optimal dilation of \( \hat{X} \). If \( \Delta \in \mathbf{D}_n(n-1) \), then \( t^*_r \geq 1 \), with equality if and only if \( \Delta \in \mathbf{D}_n(p) \).

5. **Numerical results**

We now report the results of some numerical experiments that investigated the extent to which the techniques developed in Sections 3 and 4 improve on classical MDS with respect to the raw stress (\( r = 2 \)) and stress (\( r = 1 \)) criteria. We studied the following dissimilarity matrices:

- **Voting:** This dissimilarity matrix is a disagreement matrix constructed from the voting records of \( n = 15 \) New Jersey congressmen on 19 environmental bills. Entry
$ij$ is the number of times that congressmen $i$ and $j$ voted differently. This is Data Set 294 in Hand et al. (1994), originally analyzed by Romesburg (1984).

- **Colors:** This dissimilarity matrix was constructed from Ekman’s (1954) similarity data on color perception. In Ekman’s experiment, each of 31 students rated the similarity of 91 pairs of $n = 14$ color stimuli on a scale of 0 ("no similarity at all") to 4 ("identity"). We defined corresponding dissimilarities by scaling the mean similarity ratings to range from 0 to 10, then subtracting each of them from 10.

- **Nations:** This dissimilarity matrix was constructed from Wish’s similarity data on perceptions of nations (Wish, 1971; Wish et al., 1972; Kruskal and Wish, 1978). In Wish’s experiment, each of 18 psychology students rated the similarity of 66 pairs of $n = 12$ nations on a scale of 1 ("very different") to 9 ("very similar"). The mean similarity ratings are reported in of Kruskal and Wish (1978, Fig. 7). We defined corresponding dissimilarities by subtracting each of the mean similarity ratings from 9.

- **Colas:** This dissimilarity matrix was reported by Green et al. (1989); it also appears as Table 7.1 in Groenen (1993). Each of 38 students rated the similarity of 45 pairs of $n = 10$ colas on a 9-point scale and the corresponding dissimilarities were then summed over the 38 students.

For each dissimilarity matrix $A$ and each target dimension $p = 2, 3$, we calculated the following configurations:

1. The classical solution specified in Theorem 2. This solution is denoted by $\hat{X}$.
2. The approximate solution proposed in Section 3. To obtain this solution, we first used the S-Plus function nlmnb to solve Problem (7), obtaining $\mu^+_1, \ldots, \mu^+_p$. We then proceeded as though computing the classical solution, using $\mu^+_i$ instead of $\lambda^+_i$. This solution is denoted by $X^\ast$.
3. The optimal dilations of the classical solution for the raw stress ($r = 2$) and stress ($r = 1$) criteria. These solutions are denoted by $t^*_r \hat{X}$.
4. Nominal minimizers of the raw stress ($r = 2$) and stress ($r = 1$) criteria. These were computed using PROC MDS in SAS with options fit = 2 (for $r = 2$) or fit = 1 (for $r = 1$), formula = 0, and level = absolute. These minimizers are denoted by SAS$_r$.

Instead of $\sigma_r(X)$, SAS reports

$$\sqrt{\sigma_r(X)/2} = \|D_r(X) - A_r\|_F/\sqrt{2}.$$ 

If SAS succeeds in finding a global minimizer of $\sigma_r$, then necessarily

$$\sqrt{\sigma_r(SAS)} \leq \sqrt{\sigma_r(t^*_r \hat{X})} \leq \sqrt{\sigma_r(X)}.$$ 

To quantify the error of $t^*_r \hat{X}$ relative to the error of $\hat{X}$, we computed

$$\text{RE}(t^*_r \hat{X}) = \frac{\sqrt{\sigma_r(t^*_r \hat{X})} - \sqrt{\sigma_r(SAS_r)}}{\sqrt{\sigma_r(\hat{X})} - \sqrt{\sigma_r(SAS_r)}},$$
Table 1
Relative errors (RE) for two initial configurations using the raw stress \((r = 2)\) criterion

<table>
<thead>
<tr>
<th>(\Delta)</th>
<th>(p)</th>
<th>(t_2^* \hat{X})</th>
<th>(X^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>2</td>
<td>0.4661</td>
<td>0.4637</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.6537</td>
<td>0.4714</td>
</tr>
<tr>
<td>Colors</td>
<td>2</td>
<td>0.1816</td>
<td>0.1602</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.2035</td>
<td>0.1263</td>
</tr>
<tr>
<td>Nations</td>
<td>2</td>
<td>0.4117</td>
<td>0.4023</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.3771</td>
<td>0.3023</td>
</tr>
<tr>
<td>Colas</td>
<td>2</td>
<td>0.3011</td>
<td>0.1860</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.2343</td>
<td>0.2024</td>
</tr>
</tbody>
</table>

Table 2
Relative errors (RE) for two initial configurations using the raw stress \((r = 1)\) criterion

<table>
<thead>
<tr>
<th>(\Delta)</th>
<th>(p)</th>
<th>(t_1^* \hat{X})</th>
<th>(X^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voting</td>
<td>2</td>
<td>0.7557</td>
<td>0.7820</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.8317</td>
<td>0.7339</td>
</tr>
<tr>
<td>Colors</td>
<td>2</td>
<td>0.4728</td>
<td>0.4924</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.5440</td>
<td>0.4540</td>
</tr>
<tr>
<td>Nations</td>
<td>2</td>
<td>0.2843</td>
<td>0.3795</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.3897</td>
<td>0.3969</td>
</tr>
<tr>
<td>Colas</td>
<td>2</td>
<td>0.6985</td>
<td>0.6160</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.7878</td>
<td>0.7570</td>
</tr>
</tbody>
</table>

the nominal error of the optimally dilated classical solution divided by the nominal error of the classical solution. Likewise, to quantify the error of \(X^*\) relative to the error of \(\hat{X}\), we computed the same quantity using \(X^*\) instead of \(t_2^* \hat{X}\).

Results for the raw stress \((r = 2)\) criterion are reported in Table 1. For \(r = 2\), theoretical results in Sections 3 and 4 guarantee that

\[ 0 \leq \text{RE}(X^*) \leq \text{RE}(t_2^* \hat{X}) \leq 1. \]

The numerical results suggest that both \(X^*\) and \(t_2^* \hat{X}\) substantially improve on \(\hat{X}\) in practice. In particular, the relative errors of the easily computed optimal dilations of the classical solutions ranged from 0.1816 to 0.6537.

Results for the raw stress \((r = 1)\) criterion are reported in Table 2. For \(r = 1\), theoretical results in Section 4 guarantee only that

\[ 0 \leq \text{RE}(t_1^* \hat{X}) \leq 1. \]

The numerical results suggest that both \(X^*\) and \(t_1^* \hat{X}\) improve on \(\hat{X}\) in practice. In particular, the relative errors of the easily computed optimal dilations of the classical solutions ranged from 0.2843 to 0.8317. The relative errors are somewhat larger for the raw stress criterion than for the raw stress criterion. For both criteria, however, optimal dilation of the classical solution seems well worth the modest computational expense required to obtain it.
References

Malone, S.W., Troseth, M.W., 2000b. A study of the critical points of the STRESS criterion in metric multidimensional scaling. Technical Report 00-06, Department of Computational & Applied Mathematics, Rice University, Houston, TX, USA.
