Sparsity-Exploiting Robust Multidimensional Scaling

Pedro A. Forero, Student Member, IEEE, and Georgios B. Giannakis, Fellow, IEEE

Abstract—Multidimensional scaling (MDS) seeks an embedding of $N$ objects in a $p < N$ dimensional space such that inter-vector distances approximate pair-wise object dissimilarities. Despite their popularity, MDS algorithms are sensitive to outliers, yielding grossly erroneous embeddings even if few outliers contaminate the available dissimilarities. This work introduces robust MDS approaches exploiting the degree of sparsity in the outliers present. Links with compressive sampling lead to robust MDS solvers capable of coping with unstructured and structured outliers. The novel algorithms rely on a majorization-minimization approach to minimize a regularized stress function, whereby iterative MDS solvers involving Lasso and sparse group-Lasso operators are obtained. The resulting schemes identify outliers and obtain the desired embedding at computational cost comparable to that of their non-robust MDS alternatives. The robust structured MDS algorithm considers outliers introduced by a sparse set of objects. In this case, two types of sparsity are exploited: (i) sparsity of outliers in the dissimilarities; and (ii) sparsity of the objects introducing outliers. Numerical tests on synthetic and real datasets illustrate the merits of the proposed algorithms.

Index Terms—(Block) coordinate descent, (group) Lasso, multidimensional scaling, robustness, sparsity.

I. INTRODUCTION

Multidimensional scaling (MDS) broadly refers to exploratory data tools that find an embedding (a.k.a. configuration) of $N$ objects in a $p$-dimensional vector space. The embedding is chosen such that inter-vector distances approximate the given pair-wise dissimilarities among the $N$ objects, see e.g., [9], [2]. Originally, MDS was developed in psychology to visualize via two-dimensional maps perceptual relationships among objects [32], [21]. Early applications of MDS in marketing aimed to position products in a perceptual map, successfully applied to areas ranging from high-dimensional data visualization to sensor network localization [3], [8].

Classical MDS uses the principal components of the double-centered Euclidean distance matrix to obtain the embedding when dissimilarities correspond to Euclidean distances [32]. Although able to perform well with exact distances, even a single “inconsistent” distance, hereafter termed outlier, can render the classical MDS solution of limited use. In particular, double centering spreads the effect of an outlier to the entire double-centered distance matrix, see e.g., [6]. An alternative to MDS, which can accommodate transformations on the dissimilarities as well as missing data, relies on stress functions [2, Ch. 11]. A popular algorithm to minimize the so-called raw stress is “scaling by majorizing a complicated function,” which is abbreviated as SMACOF [10]. The stress function is a weighted sum of squared-errors between dissimilarities and embedding inter-vector distances. Unfortunately, the dependency of stress functions on the least-squares (LS) criterion renders SMACOF and related MDS schemes sensitive to outliers [17].

Despite the popularity of MDS in various applications, dealing with outliers has been relegated to pre- and post-processing tasks, which remove “suspicious” dissimilarities based on “large” residual errors [9]. These two-step approaches to find the embedding and remove outliers have to decide when it is justified to discard dissimilarities. One of the first systematic approaches to tackle the outlier sensitivity of MDS resorted to tools from robust statistics for dissimilarities expressed as Euclidean distances [28]. The embedding was obtained by heuristically modified Newton-Raphson iterations solving a nonlinear system of equations. A criterion which exploited the known resilience to outliers of the $\ell_1$-error norm was proposed to obtain a robust Euclidean embedding (REE) [6]. The resulting solvers were based on semidefinite programming and sub-gradient descent methods. Although rooted on a robust criterion, it was empirically observed that REE yields embeddings that underestimate the given dissimilarities, thus it further requires a robust scale estimate to properly adjust the embedding. In the context of sensor networks, the Huber function was employed to find an embedding corresponding to sensor positions [20]. The resulting algorithm used a two-level majorization-minimization (MM) strategy, where each minimization step of the first level iteration was solved through an MM algorithm. As acknowledged in [20], this nested MM structure slows down convergence.

The motif behind the robust MDS algorithms developed in this work is the degree of sparsity present in the out-
liers contaminating the dissimilarity data. Leveraging this knowledge, robust MDS based on the least-trimmed squares (LTS) criterion is proposed in Section II. It is shown that robust MDS via LTS is equivalent to a regularized LS approach stemming from a dissimilarity model that explicitly accounts for outliers. The regularization term corresponds to the $\ell_0$-norm and establishes a link between robust MDS and the area of compressive sampling [4], [33]. Capitalizing on this link, the $\ell_0$-norm is relaxed to its closest convex approximation, namely the $\ell_1$-norm. An MM-based iterative algorithm is developed in Section III, which involves closed-form solutions of scalar least-absolute shrinkage and selection operator (Lasso) problems. The resulting algorithm exhibits computational complexity comparable to that of SMACOF. In practice, outliers can appear due to just a few faulty objects. In this case, objects introducing outliers are also considered to be sparse. Moreover, these faulty objects induce grouping of the outliers. A robust structured MDS algorithm seeking a few sparse groups of outliers is developed in Section IV. The resulting algorithm features closed-form updates corresponding to group Lasso solvers [34]. Numerical tests to illustrate the performance of the proposed robust MDS algorithms on both synthetic and real data are presented in Section V. Concluding remarks are given in Section VI.

From a high-level viewpoint, the main contributions of this work are: (i) principled robust MDS approaches based on a universal criterion linking robust statistics with compressive sampling; (ii) provably convergent iterative algorithms with closed-form updates and relatively low computational complexity; and, (iii) a versatile MDS framework, which can accommodate missing data and structured outliers.

Notational conventions are as follows: upper (lower) bold face letters are used for matrices (column vectors); $(\cdot)^T$ denotes transposition; $[\cdot]_{n,m}$ the $(n,m)$-entry of a matrix (n-entry of a vector); $\text{Tr}(X)$ the trace operator of matrix $X$; $\| \cdot \|_2$ the Euclidean (Frobenious) norm; $\{X^{(t)}\}_{t \geq 1}$ the sequence $X^{(1)},X^{(2)},X^{(3)},\ldots$; $0_{N \times N}$ the $N \times N$ matrix of zeros; $1_N (0_N)$ the $N \times 1$ vector of ones (zeros); $S^N$ the set of $N \times N$ symmetric matrices; $S^N_+$ the set of $N \times N$ symmetric positive matrices; $S^N_-$ the set of $N \times N$ symmetric skew matrices; $S^N_h$ the set of $N \times N$ symmetric hollow matrices (with zeros on the main diagonal), where $\Delta \in S^N_h$ if and only if $\Delta \in S^N$ and $[\Delta]_{n,n} = 0$, $\forall n \in \{1, \ldots, N\}$.

II. BACKGROUND AND PROBLEM FORMULATION

Consider the set of pairwise dissimilarities $D := \{\{d_{n,m}\}_{n=1}^{N} \}_{m=n+1}^{N}$ between $N$ objects, where $d_{n,m}$ denotes the real-valued dissimilarity between objects $n,m \in \{1, \ldots, N\}$. The dissimilarities satisfy: (i) $d_{n,m} \geq 0$; (ii) $d_{n,m} = d_{m,n}$; and, (iii) $d_{n,n} = 0$ for all $n$. Hence, the total number of distinct pairwise dissimilarities for $N$ objects is $N_0 := N(N-1)/2$. As $d_{n,m}$ are not necessarily distances, they are not required to satisfy the triangle inequality. Given $D$, MDS seeks an embedding of $N$ vectors in a $p$-dimensional space, where the embedding dimension $p < N$ is prescribed. The embedding is given by the columns of $X := [x_1, \ldots, x_N] \in \mathbb{R}^{p \times N}$, which is chosen so that pairwise vector distances $d_{n,m}(X) := \|x_n - x_m\|_2$ approximate $d_{n,m}$ in some well defined sense.

Aiming to account for outliers, let $\{r_{\rho}(X)\}_{\rho=1}^{\rho_0}$ denote ordered residuals $r_{\rho}(X) := d_{n,m} - d_{n,m}(X)$; where $\rho := (n-1)N - n(n+1)/2 + m$; that is, $r_{(1)}(X) \leq r_{(2)}(X) \leq \cdots \leq r_{(\rho_0)}(X)$. For a given trimming constant $R \leq \rho_0$, the LTS estimate of $X$ is given by the solution of [27]

$$\min_X \sum_{\rho=1}^{R} r_{\rho}(X), \quad (1)$$

Given $X$ and $R$, the cost in (1) can be computed by: (i) evaluating the squared residuals $\{r_{\rho}(X)\}_{\rho=1}^{\rho_0}$; (ii) sorting them in ascending order to obtain $\{r_{(\rho)}(X)\}_{\rho=1}^{\rho_0}$; and, (iii) adding the smallest $R$ residuals. Let $R \subset \rho_0$ denote any subset of dissimilarities with cardinality $|R| = R$. Then, (1) can be written as

$$\min_X \min_{R: |R| = R} \sum_{\rho \in R} [d_{n,m} - d_{n,m}(X)]^2. \quad (2)$$

In order to solve (2) one has to consider all $C := \binom{\rho_0}{R}$ subsets $\{R_c\}_{c=1}^{C}$ with cardinality $R$. For each $R_c$, the outer minimization in (2) can be performed by solving a nonlinear LS problem. After solving (2) for each $R_c$, the embedding estimate $\hat{X}_c$ and its corresponding cost $\sigma(\hat{X}_c) := \sum_{d_{n,m} \in \rho_c} (d_{n,m} - d_{n,m}(X))^2$ can be obtained. Finally, the minimizer of (1) is set to $\hat{X} = \hat{X}_{c^*}$, where $c^* = \arg\min_c \sigma(\hat{X}_c)$.

Problem (2) is combinatorial in nature and its solution entails solving $C$ nonlinear LS subproblems. Thus, its computational complexity renders (2) impractical even for small values of $N$. Moreover, each nonlinear LS problem is nonconvex, thus global optimality of $\hat{X}$ cannot be guaranteed. Note that solving (2) is tantamount to performing hard rejection of outliers prior to estimating $X$.

A. Outlier-aware dissimilarity model

A robust MDS approach capitalizing on a dissimilarity model that explicitly accounts for outliers is developed in this section. To this end, an outlier variable $\delta_{n,m}$ is introduced per $d_{n,m}$. An $\mathbf{o}_{n,m} \neq 0$ means that $\delta_{n,m}$ is deemed as outlier; otherwise, $\mathbf{o}_{n,m}$ is set to zero and $\delta_{n,m}$ is deemed as belonging to the nominal (outlier-free) set of dissimilarities. The advocated dissimilarity model is thus

$$d_{n,m} = d_{n,m}(X) + \mathbf{o}_{n,m} + \epsilon_{n,m}, \quad \forall n \in \{1, \ldots, N\}, \ m \geq n+1$$

where $\{\epsilon_{n,m}\}$ denote zero-mean independent random variables modeling nominal errors. Variables $\{\mathbf{o}_{n,m}\}$ are collected in the outlier matrix $O \in S^N_h$ whose entries are $[O]_{n,m} := \mathbf{o}_{n,m}$.

Even when all $\{\mathbf{o}_{n,m}\}$ are available, solving for $O$ and $X$ in (3) is an under-determined problem entailing $N_0 + pN$ unknowns in $N_0$ equations. However, outliers are typically rare and, presumably only few of them appear in $D$. The scarcity of outliers contaminating $D$ translates to sparsity in the entries of $O$. This observation is instrumental to develop a robust algorithm for estimating both $O$ and $X$.

Consider the indicator function $1(\mathbf{o}_{n,m} \neq 0) := 1$, if $\mathbf{o}_{n,m} \neq 0$; and $1(\mathbf{o}_{n,m} \neq 0) := 0$, otherwise; and also
the shorthand notation \( \sum_{n<m} \delta_{n,m} := \sum_{n=1}^{N} \sum_{m=n+1}^{N} \delta_{n,m} \).

Then, an estimator for \((\hat{O}, \hat{X})\) is

\[
(\hat{O}, \hat{X}) = \arg \min_{O \in \mathbb{R}^N, X} \sum_{n<m} [\delta_{n,m} - d_{n,m}(X) - o_{n,m}]^2 + \lambda \sum_{n<m} 1(o_{n,m} \neq 0) \tag{4}
\]

where the tuning parameter \( \lambda > 0 \) regulates the sparsity of \( O \). The regularization term in (4) corresponds to the \( \ell_0 \)-norm of \( O \). If \( X \) were known, the cost in (4) would decouple across each \((n,m)\)-pair and problem (4) would reduce to \( N \) scalar optimization problems to obtain all outlier variables. Finding \((\hat{O}, \hat{X})\) requires solving \( 2^N \) nonlinear LS problems when no extra information is available. This exponential complexity can be reduced by considering side information on the degree of sparsity (number of outliers) \( S := N_0 - R \) present in \( O \).

To gain insight on the form of the minimizer \((\hat{O}, \hat{X})\), suppose that \( \lambda = \lambda_S \) is chosen so that \( \sum_{n<m} 1(\hat{o}_{n,m} \neq 0) = S \). Then, the minimizer of (4) with respect to \( O \) is given in closed form for each non-trivial entry of \( O \) as (recall that \( \rho := (n-1)N - n(n+1)/2 + m \))

\[
\hat{o}_{n,m} = \begin{cases} 
0 & \text{if } r^2_\rho(X) \leq \lambda_S \\
\frac{r^2_\rho(X)}{r^2_\rho(X) > \lambda_S} & \text{otherwise}
\end{cases} \tag{5}
\]

The form of (5) shows that the effect of \( \hat{o}_{n,m} \) in (4) is to remove the largest \( S \) residuals from the cost.

If the number of outliers \( S \) were known a priori, then solving (4) for \( X \) would entail solving \( C = (N^2_S)^{-1} \) nonlinear LS problems. To this end, one must consider the collection of sets \( \{D_c \subset D\}_{c=1}^{C} \) satisfying \( |D_c| = S \). Then, using (5) in (4) and the set differences \( R_c := D - D_c \), it is possible to find the collection of embeddings \( \{X_c\}_{c=1}^{C} \) as

\[
\hat{X}_c = \arg \min_{\delta_{n,m} \in R_c} \sum_{n<m} |\delta_{n,m} - d_{n,m}(X)|^2, \quad c = 1, \ldots, C. \tag{6}
\]

Finally, \( \hat{X} \) is chosen from \( \{\hat{X}_c\}_{c=1}^{C} \) as the one returning the smallest cost in (6). The previous discussion establishes the following result.

**Proposition 1.** If \( \lambda \) is chosen so that \( \sum_{n<m} 1(\hat{o}_{n,m} \neq 0) = S \), and \((\hat{X}, \hat{O})\) is a local minimum for (4), then \( \hat{X} \) is also a local minimum for (2) with \( R = N_0 - S \).

Proposition 1 justifies the model in (3) and the regularized problem (4) by establishing its equivalence with the LTS criterion [27]. It is important to stress that the solution of (4) is generally non-unique. At the onset, finding \( \hat{X} \) in (4) even with \( \hat{O} \) given, requires a nonconvex (and thus non-efficient) solver. Clearly, rigid motions of \( X \) including translations, rotations and reflections leave inter-vector distances unaltered, and thus are indistinguishable for the cost in (4). Hence, each local optimum of (4) specifies a family of embeddings, via rigid motions, yielding the same \( \{d_{n,m}(X)\} \).

The estimator in (4) discards all dissimilarities deemed as outliers; thus, it is considered as a hard outlier-rejection strategy. On the other hand, soft outlier-rejection strategies for robust estimation are often favored since they can include potentially useful information to the estimator [17]. Moreover, as shown in the ensuing sections, they can lead to solvers of reduced complexity. In the next section, a soft outlier-rejection algorithm for MDS stemming from a convex relaxation of the \( \ell_0 \)-norm is developed.

### B. Outlier rejection via \( \ell_1 \)-norm regularization

In order to obtain computationally tractable alternatives to (4), the option pursued here is a surrogate cost function. Recalling that \( |o_{n,m}| \) is the closest convex approximation to \( 1(o_{n,m} \neq 0) \) [33], compressive sampling and related sparsity-aware signal processing methods [33], [4], suggest surrogating the cost in (4) by what can be termed \( \ell_1 \)-norm regularized stress

\[
f(O, X) := \sum_{n<m} [\delta_{n,m} - d_{n,m}(X) - o_{n,m}]^2 + \lambda \sum_{n<m} |o_{n,m}| \tag{7}
\]

and estimating \((O, X)\) by solving

\[
\min_{O \in \mathbb{R}^N, X} f(O, X). \tag{8}
\]

The first term of \( f(O, X) \) measures the goodness-of-fit of pairwise distances among columns of \( X \) with those in \( D \) after the influence of the outliers has been removed, whereas the second term is a regularizer capturing the degree of sparsity in \( O \). The tuning parameter \( \lambda \) affects the number of dissimilarities deemed as outliers. If \( \lambda \rightarrow \infty \), then all \( \hat{o}_{n,m} = 0 \) and (8) boils down to MDS which minimizes the so-called raw stress [2, Ch. 3]. If \( \lambda \rightarrow 0^+ \), then all \( \hat{o}_{n,m} \) are non-zero. This choice of \( \lambda \) can yield degenerate solutions. For instance, setting \( \hat{o}_{n,m} = \delta_{n,m} - d_{n,m}(X) \) drives the first term in (7) to zero for any \( X \). Then, setting \( x_n = 0, \forall n \in \{1, \ldots, N\} \), and \( \hat{o}_{n,m} = \delta_{n,m} \) yields the minimum \( \ell_1 \)-norm regularized stress value zero.

A computationally efficient solver for (8) using MM is developed in the ensuing section. But before that, a couple of remarks are in order.

**Remark 1.** (Connection to robust M-estimators) In the context of linear regression and under the assumption that nominal noise plus outliers follow an \( \epsilon \)-contaminated distribution [17], a problem related to (8) has been shown equivalent to one adopting Huber’s function of the residuals [13]. The equivalence and performance claims for this M-estimator hold for a specific choice of \( \lambda = 2H \), where \( H \) is the parameter of the Huber function. For a linear regression model with additive Gaussian noise, setting \( H = 1.345\sigma \), where \( \sigma \) is the standard deviation of the nominal errors, yields an estimator with 95% efficiency [17]. Contrastingly, the outliers in (3) are not tied to a specific statistical model, and \( \lambda \) is selected from the data. From this vantage point, (8) is a universal criterion able to accommodate general outlier models through a judicious selection of the regularization term [14].

**Remark 2.** (Missing data and weighted residuals) In marketing and sensor localization applications some \( \delta_{n,m} \) are not available. These missing dissimilarities can be accounted in the present framework through binary \((0,1)\) weights \( w_{n,m} \) after
redefining (7) as
\[ f_w(O, X) := \sum_{n<m} w_{n,m} [\delta_{n,m} - d_{n,m}(X) - o_{n,m}]^2 + \lambda \sum_{n<m} w_{n,m} |o_{n,m}| \tag{9} \]
and estimating \((O, X)\) by solving
\[ \min_{O \in \mathbb{S}_h^N} f_w(O, X) \tag{10} \]
where \(w_{n,m} = 1\) if \(\delta_{n,m}\) is available, and \(w_{n,m} = 0\) otherwise. In general, however, any real \(w_{n,m} > 0\) is allowed if \(\delta_{n,m}\) is available. Sammon mapping, for instance, uses \(w_{n,m} = \delta_{n,m}^{-1}\) to de-emphasize large dissimilarities [19]. Other strategies to select \(w_{n,m}\) can be found in [2, Ch. 11].

III. ROBUST MDS VIA MAJORIZATION

In this section, an iterative solver is developed for (8) using the MM approach. A brief description of MM is given first (Section III-A). Then, a majorizer of \(f\) in (7) is introduced (Section III-B), to yield an iterative algorithm for solving (8) (Section III-C).

A. Majorization-Minimization algorithm

The MM approach relies on a real-valued function \(g\), called majorizer, which satisfies
\[ g(O, X; Q, Z) \geq f(O, X) \tag{11a} \]
\[ g(O, X; O, X) = f(O, X) \tag{11b} \]
where \(Q \in \mathbb{S}_h^N\) and \(Z := [z_1, \ldots, z_N] \in \mathbb{R}^{P \times N}\) are both auxiliary matrices. Condition (11a) clearly holds for those \(g\) upper bounding \(f\), \(\forall (O, X)\). Moreover, the majorizer \(g\) “touches” \(f\) when \((Q, Z) = (O, X)\) as dictated by (11b). With \(t \in \mathbb{N}\) indexing iterations, an MM algorithm amounts to
\[ (O^{(t+1)}, X^{(t+1)}) = \arg \min_{O \in \mathbb{S}_h^N} g(O, X; O^{(t)}, X^{(t)}) \tag{12} \]
where \(O^{(t)}\) and \(X^{(t)}\) denote \(O\) and \(X\) estimates at iteration \(t\). The MM algorithm minimizes \(g\) with respect to \((O, X)\), redefines \(g\) using the newly updated values, and iterates until convergence. It is important to remark that success of MM hinges on selecting \(g\) so that (12) can be solved easily.

Observe that the real-valued sequence \(\{f(O^{(t)}, X^{(t)})\}\) is non-increasing. This holds because
\[ f(O^{(t)}, X^{(t)}) = g(O^{(t)}, X^{(t)}; O^{(t)}, X^{(t)}) \geq g(O^{(t+1)}, X^{(t+1)}; O^{(t)}, X^{(t)}) \geq g(O^{(t+1)}, X^{(t+1)}; O^{(t)}, X^{(t+1)}) = f(O^{(t+1)}, X^{(t+1)}) \tag{13a} \]
\[ = f(O^{(t+1)}, X^{(t+1)}) \tag{13b} \]
where (13a) holds because of (12), and (13b) follows after using (11b) into (11a). In fact, for the sequence \(\{f(O^{(t)}, X^{(t)})\}\) to be non-increasing it suffices that \((O^{(t+1)}, X^{(t+1)})\) satisfies per iteration
\[ g(O^{(t+1)}, X^{(t+1)}; O^{(t)}, X^{(t)}) \leq g(O^{(t)}, X^{(t)}; O^{(t)}, X^{(t)}). \tag{14} \]

B. Majorizing the \(\ell_1\)-norm regularized stress

In this section, a majorizer \(g\) for \(f\) in (7) is proposed. Let matrix \(\Delta \in \mathbb{S}_h^N\) be formed with entries \((\Delta_{n,m} := \delta_{n,m})\). Recall that \(f(O, X)\) in (7) is nonconvex and non-differentiable with respect to \(X\). Expanding \(f(O, X)\) yields
\[ f(O, X) = \frac{1}{2} \|\Delta - O\|^2_F - 2 \sum_{n<m} (\delta_{n,m} - o_{n,m}) d_{n,m}(X) + \sum_{n<m} d_{n,m}(X) + \lambda \frac{1}{2} \|O\|_1 \tag{15} \]
where \(\|O\|_1 := \sum_{n<m} |o_{n,m}|\). Let \(\{e_n \in \mathbb{R}^N\}_{n=1}^N\) denote the set indicator vectors with entries \(e_n[m] = 1\) if \(m = n\), and zero if \(m \neq n\). The term \(\sum_{n<m} d_{n,m}(X)\) can be written in compact form as
\[ \sum_{n<m} d_{n,m}(X) = \sum_{n<m} (e_n - e_m)^T X^T X (e_n - e_m) = \text{Tr}(XLX^T) \tag{16} \]
where \(L \in \mathbb{S}_h^N\) has off-diagonal entries \([L]_{n,m} = -1\) and main diagonal entries \([L]_{n,n} = N - 1\) [2, Ch. 8]. Matrix \(L\) corresponds to the Laplacian matrix of the complete graph \(G(V,E)\), whose vertices correspond to columns of \(X\), that is, \(V := \{x_1, \ldots, x_N\}\), and all edges \(E\) have unit weights.

The second term in (15) is non-differentiable and nonconvex. Specifically, if \(\delta_{n,m} > o_{n,m}\) for some \(n, m \in \{1, \ldots, N\}\), then \(-2 \sum_{n<m} (\delta_{n,m} - o_{n,m}) d_{n,m}(X)\) is nonconvex. The ensuing Proposition 2 (proved in Appendix A) will define a majorizer for \(-2 \sum_{n<m} (\delta_{n,m} - o_{n,m}) d_{n,m}(X)\) that is differentiable with respect to both \(X\) and \(O\). Differentiability with respect to \(X\) is instrumental to develop a solver for (8) with affordable computational complexity.

Before specifying the majorizer, let \(P(O, Z) := \{(n,m) : \delta_{n,m} > o_{n,m}, d_{n,m}(Z) > 0\}, P'(O, Z) := \{(n,m) : \delta_{n,m} \leq o_{n,m}, d_{n,m}(Z) > 0\}\) denote pair-index sets, and define matrices \(A_1(O, Z), A_2(O, Z) \in \mathbb{S}_h^N\) with entries
\[ [A_1(O, Z)]_{n,m} := \begin{cases} \frac{\delta_{n,m} - o_{n,m}}{d_{n,m}(Z)} & (n,m) \in P(O, Z) \\ 0 & \text{otherwise} \end{cases} \tag{17a} \]
\[ [A_2(O, Z)]_{n,m} := \begin{cases} -\frac{\delta_{n,m} - o_{n,m}}{d_{n,m}(Z)} & (n,m) \in P'(O, Z) \\ 0 & \text{otherwise} \end{cases} \tag{17b} \]
Matrices \(A_1\) and \(A_2\) are adjacency matrices corresponding to weighted graphs \(G_1(V, E_1)\) and \(G_2(V, E_2)\), respectively, where \(E_1\) and \(E_2\) are defined through the nonzero entries of \(A_1\) and \(A_2\). A weighted edge \((n,m)\), with weight \(|\delta_{n,m} - o_{n,m}| / d_{n,m}(Z)_n\), between the nodes corresponding to \(x_n\) and \(x_m\) in \(G_1\) (\(G_2\)) exists if and only if \((n,m) \in P(O, Z)\) (resp. \((n,m) \in P'(O, Z)\)). The corresponding Laplacian matrices for \(G_1\) and \(G_2\) are given by
\[ L_1(O, Z) = \text{diag}(A_1(O, Z)1_N) - A_1(O, Z) \tag{18a} \]
\[ L_2(O, Z) = \text{diag}(A_2(O, Z)1_N) - A_2(O, Z) \tag{18b} \]
Proposition 2. A majorizer of \(-2 \sum_{n<m} (\delta_{n,m} - o_{n,m})d_{n,m}(X)\) in (15) is given by

\[
g_0(O, X; Z) = -2\text{Tr} (XL_1(O, Z)Z^T) + \text{Tr} (XL_2(O, Z)X^T) + C_0(O, Z)
\]

where \(C_0(O, Z) := -\sum_{n<m} (\delta_{n,m} - o_{n,m})d_{n,m}(Z)\).

Substituting (16) into (15) and using the majorizer \(g_0\) yields the desired majorizer \(g\) as

\[
g(O, X; Z) = \text{Tr} (X(L + L_2(O, Z))X^T) + C_0(O, Z) \quad \text{where} \quad g(O, X; Z) \geq f(O, X), \forall (O, X), \text{with equality holding if } Z = X.
\]

The majorizers in (19) and (20) have been dropped from their argument.

C. Robust MDS via generalized MM

The MM iteration to solve (8) is [cf. (12)]

\[
(O^{(t+1)}, X^{(t+1)}) = \arg \min_{O \in \bar{S}^N, X} g(O, X; X^{(t)})
\]

which requires jointly minimizing \(g\) with respect to \(O\) and \(X\). However, \(g\) in (20) is nonconvex with respect to \((O, X)\) due to the products \(XL_1(O, Z)\) and \(XL_2(O, Z)X^T\), as well as non-differentiable with respect to \(O\) due to \(\|O\|_1\). Hence, solving (21) may incur prohibitively high complexity per iteration. To alleviate this, the idea is to capitalize on the weaker condition (14), and replace (21) by a pair of updates that can be expressed in closed form.

The desired majorizer \(g\) is convex with respect to each optimization variable \(O\) and \(X\), and differentiable with respect to \(X\). This fact suggests replacing (21) by a sequence of minimization steps, one per optimization variable, as

\[
O^{(t+1)} = \arg \min_{O \in \bar{S}^N} g(O, X^{(t)}; X^{(t)}) \quad \text{and} \quad X^{(t+1)} = \arg \min_{X} g(O^{(t+1)}, X; X^{(t)}).
\]

Since (22a) and (22b) satisfy (14), they can be used in lieu of (21) to devise an MM algorithm for solving (8). Next, it is shown that the updates in (22) are solvable in closed form.

First consider (22a) and recall that since \(g\) is a majorizer satisfying (11b), the cost in (22a) corresponds to \(f(O, X^{(t)})\). Then, update (22a) is equivalent to

\[
O^{(t+1)} = \arg \min_{O \in \bar{S}^N} f(O, X^{(t)})
\]

which entails solving \(N_0\) scalar sub-problems as shown next. Indeed, each \(o_{n,m}^{(t+1)}\) entry corresponds to the solution of

\[
\min_{o_{n,m}} (\delta_{n,m} - d_{n,m}(X^{(t)}) - o_{n,m})^2 + \lambda |o_{n,m}|
\]

which is a scalar Lasso problem whose solution is expressible using the soft-thresholding operator \(S_{\lambda}'(x) := \text{sgn}(x)(|x| - \lambda/2)_+\), with \(\text{sgn}(\cdot)\) denoting the sign function, and \((\cdot)_+ := \max\{0, \cdot\}\) [31]. Using the latter, the solution of (24) becomes available in closed form as

\[
o_{n,m}^{(t+1)} = S_{\lambda}(\delta_{n,m} - d_{n,m}(X^{(t)})).
\]

The functional form of (25) illustrates how \(\lambda\) affects the number of \(\{\delta_{n,m}\}\) dissimilarities deemed as outliers. The effect of \(\lambda\) is to zero the \(\{o_{n,m}^{(t+1)}\}\) whose corresponding absolute residuals \(\{|\delta_{n,m} - d_{n,m}(X^{(t)})|\}\) are less than \(\lambda/2\). Dissimilarities \(\delta_{n,m}\) whose corresponding \(o_{n,m}^{(t+1)} \neq 0\) are considered outliers, and their effect on (7) is tapered through \(o_{n,m}^{(t+1)}\).

A useful property of (25) is that \(\delta_{n,m} > o_{n,m}^{(t+1)}\) for all iterations \(t\). This holds trivially whenever \(o_{n,m}^{(t+1)} \leq 0\) since \(\delta_{n,m} > 0\). If \(o_{n,m}^{(t+1)} > 0\), then one must show that \(\delta_{n,m} - o_{n,m}^{(t+1)} > 0\). Using (25) in the previous inequality leads to

\[
\delta_{n,m} - o_{n,m}^{(t+1)} = \delta_{n,m} - \left(\delta_{n,m} - d_{n,m}(X^{(t)}) - \lambda\right)
\]

\[
= d_{n,m}(X^{(t)}) + \frac{\lambda}{2} > 0
\]

where the first equality follows since \(o_{n,m}^{(t+1)} > 0\), and the inequality since \(d_{n,m}(X^{(t)}) \geq 0\) and \(\lambda > 0\). The following lemma summarizes the aforementioned property.

Lemma 1. For all iterations \(t\), the outlier estimate in (25) satisfies \(o_{n,m}^{(t+1)} < \delta_{n,m}\).

Updating \(X^{(t+1)}\) via (22b) boils down to solving a quadratic program. The majorizer \(g\) is defined through matrices \(L_1(O^{(t+1)}, X^{(t)})\) and \(L_2(O^{(t+1)}, X^{(t)})\). These matrices are defined as in (18), through the now iteration-dependent sets \(P^{(t+1)} = (\{n,m\} : \delta_{n,m}^{(t+1)} > o_{n,m}^{(t+1)}, d_{n,m}(X^{(t)}) > 0\) and \(P'\) as \(\{n,m\} : \delta_{n,m}^{(t+1)} < o_{n,m}^{(t+1)}, d_{n,m}(X^{(t)}) > 0\). In particular, Lemma 1 implies that \(P'\) remains empty for all iterations; thus, \(L_2(O^{(t+1)}, X^{(t)})\) is always zero. This fact together with the first-order optimality condition for (22b) yield the closed-form solution

\[
X^{(t+1)} = X^{(t)} L_1(O^{(t+1)}, X^{(t)}) L_1^{\dagger}
\]

where \((\cdot)^\dagger\) denotes the Moore-Penrose pseudoinverse. Using \(L_1^{\dagger}\) yields the minimum Frobenious-norm solution for \(X^{(t+1)}\).

The minimum-norm solution removes the translation ambiguity with respect to \(X\) inherent in (20) by forcing \(X^{(t+1)}\) to be centered at the origin, that is, \(X^{(t+1)} 1_N = 0\) for all \(t\).

Next, the convergence of the updates in (22) is considered. Toward this goal, define a pair \((O, X)\) usable if \(d_{n,m}(X) > 0\) for \(\delta_{n,m} \neq o_{n,m}, \forall n, m\). With \((O, X)\) denoting a limit point of the sequence \(\{(O_{i}, X)_{i}^{(t)}\}_{t \geq 1}\) generated by (22), the following result characterizes the behavior of \(\{(O_{i}, X_{i})_{i}^{(t)}\}_{t \geq 1}\) and its set of limit points \(\bar{\Omega}\) (see Appendix B for the proof).

Proposition 3. The limit points \((O, X) \in \bar{\Omega}\) of (22a) and (22b) are usable local minima for (7). Moreover, the set \(\bar{\Omega}\) is connected.

Proposition 3 implies that the limit points of (22) take values in a continuum for each local minimum of (7). A natural
Algorithm 1 Robust MDS

Require: Set $O^{(0)}$ to zero and randomly initialize $X^{(0)}$.
1: for $t = 0, 1, 2, \ldots$ do
2: \hspace{1em} Find each entry of $O^{(t+1)}$ via (25).
3: \hspace{1em} Update $X^{(t+1)}$ via (27) with $L_1$ as in (18a) and (17a).
4: end for

candidate for such a continuum is \{($\tilde{O}, X) : X = V X$\}, where $V \in \mathbb{R}^{p \times p}$ denotes an orthogonal matrix. The following proposition establishes the convergence of \{f$(O^{(t)}, X^{(t)})$\}$t \geq 1$(see Appendix C for the proof).

Proposition 4. The sequence \{f$(O^{(t)}, X^{(t)})$\}$t \geq 1$ generated by RMDS converges to f$(O, X)$.

Although \{($O^{(t)}, X^{(t)})$\}$t \geq 1$ can have an infinite number of limit points, Proposition 4 holds since (7) is invariant to rotations with respect to $X$.

The robust MDS (RMDS) scheme is summarized as Algorithm 1. RMDS terminates when the relative error
\[ \|X^{(t+1)} - X^{(t)}\|_F/\|X^{(t)}\|_F \] becomes smaller than a small positive threshold $\nu$, e.g., $\nu = 10^{-6}$. It is argued next that the computational complexity and storage requirements of RMDS are of the same order as the ones for the non-robust SMACOF algorithm.

Remark 3. (Time and space complexity) Per iteration RMDS updates $O$ entry-wise through (25). This involves solving $N_0$ scalar Lasso problems, each incurring complexity $O(p)$ due to the Euclidean distance update. Updating $X$ via (27) requires computing two matrix multiplications with $O(N^3)$ complexity in the worst case. The $N \times N$ matrix $L_1$ is computed off-line before the algorithm starts, entailing $O(N^3)$ computations. Moreover, if all $w_{n,m}$ are available and all $w_{n,m} = 1$, then $L_1 = N^{-1}(I - N^{-1}1_N1_N^T)$, which incurs $O(N^2)$ complexity. Matrix $L_1$ is also updated per iteration, incurring $O(N^2)$ operations. Hence, the total computational complexity per iteration is $O(N^3)$. The space requirements of RMDS are: (i) a $p \times N$ matrix $X$ containing $Np$ scalars; and, (ii) three $N \times N$ symmetric matrices, namely, $O$, $L$, and $L_1$, requiring storage of $N(N+1)/2$ scalars each. Thus, RMDS requires $O(N^2)$ memory. Since $O$ is sparse, it can be stored efficiently by using sparse matrix structures.

Remark 4. (Selection of $\lambda$) Choosing $\lambda$ controls the number of dissimilarities deemed as outliers, thereby affecting the resultant embedding. Whenever side information is available, $\lambda$ can be chosen through a grid search over a set of values \{$\lambda_g$\}$g=1$. RMDS is run for each $\lambda_g$ yielding solution pairs $(X_g, \tilde{O}_g)$. When the number of outliers $S$ is known, then one can select $\lambda = \lambda_g$ so that $\sum_{n,m \in T_g} |\tilde{O}_{g,n,m}| = S$. Instead, if the nominal noise variance $\sigma^2$ is available, then it is possible to choose $\lambda$ based on a sample estimate of $\sigma^2$ obtained from $D$ after removing the outliers. Let $T_g := \{(n,m) : |\tilde{O}_{g,n,m}| = 0\}$ denote the set of dissimilarities deemed as outlier-free per $\lambda_g$. Then, an estimate of the noise variance $\hat{\sigma}^2$ is obtained as
\[ \hat{\sigma}^2_g = \frac{1}{|T_g|} \sum_{(n,m) \in T_g} [d_{n,m} - d_{n,m}(\tilde{X}_g)]^2 \] (28)

and $\lambda = \lambda_g^*$ for $g^* = \arg \min_g (\sigma^2 - \hat{\sigma}^2)^2$.

This collection of RMDS problems can be solved efficiently by sorting the grid points $\lambda_g$ in a decreasing order $\lambda_{max} > \cdots > \lambda_g = \lambda_{min} > 0$, running RMDS for each $\lambda_g$ from $\lambda_{max}$ to $\lambda_{min}$, and using warm starts to define the initialization of each successive RMDS run. Here, warm starts refer to initializing RMDS for $\lambda_g$ with $(X_{g-1}, \tilde{O}_{g-1})$ obtained by RMDS for the previous $\lambda_g$ in the sequence. A similar grid search with warm starts has been demonstrated to exhibit fast convergence rates in the context of (block) coordinate descent methods for Lasso and group Lasso solvers [12].

IV. ROBUST MDS FOR STRUCTURED OUTLIERS

A robust approach for MDS, able to handle structured outliers contaminating $D$ is developed in this section. The specific structure considered assumes that outliers are present in $D$ due to a small group of “faulty” objects. This situation arises, for example, in sensor localization problems where few faulty sensors generate erroneous inter-sensor distances. In this case, outliers appear grouped in the rows and columns of $\Delta$ corresponding to the faulty sensors.

This section begins by proposing an appropriate regularized stress function, termed $\Gamma$-regularized stress, to account for the outlier group structure (Section IV-A). Then, a suitable majorizer for the $\Gamma$-regularized stress is proposed (Section IV-B), leading to the development of an MM-based solver (Section IV-C).

A. Sparse-group regularization for structured outliers

Even with structured outliers, the goal of MDS remains to find a $p$-dimensional embedding $X$ based on the dissimilarity matrix $\Delta$, where dissimilarities abide by the model (3). However, it is now assumed that only a few objects introduce outliers in $\Delta$. In this case, it is convenient to replace $O$ by $B + C$, where $B, C \in \mathbb{R}^{N \times N}$ are auxiliary matrices with zeros on their main diagonals. Since $\Delta$ is a symmetric matrix, $B + C$ should be symmetric too. One way to ensure this is to have $B = CT$ in which case $B + C = CT + C$ is indeed symmetric.

Now one seeks $B := [b_1, \ldots, b_N], C := [c_1, \ldots, c_N]^T$, and $X$ as the minimizers of
\[ \min_{B,C,X} f_\Gamma(B, C, X) \] (29)
where the $\Gamma$-regularizer stress $f_\Gamma$ is defined as
\[ f_\Gamma(B, C, X) := \frac{1}{2} \|\Delta - D(X) - (B + C)\|_F^2 \] (30)
\[ + \Gamma(B, C) + \frac{\mu}{2} \|B - CT\|_F^2 \]
$D(X) \in S^N_{b}$ is an Euclidean distance matrix with entries $D(X)_{n,m} := d_{n,m}(X)$, and $\mu > 0$ is a tuning parameter. The term $\mu\|B - CT\|_F^2$ penalizes disagreement between $B$ and $CT$, while $\Gamma$ is chosen to induce the desired sparse group structure on the outliers, and is given by $(\gamma, \lambda > 0$ are tuning parameters)
\[ \Gamma(B, C) := \gamma \sum_{n=1}^{N} (\|b_n\|_2 + \|c_n\|_2) + \lambda \sum_{n=1}^{N} (\|b_n\|_1 + \|c_n\|_1). \] (31)
The first sum in (31) penalizes the columns of \( B \) and rows of \( C \) through their corresponding Euclidean norms, which are known to induce sparsity of groups of variables [34]. In particular, \( \|b_n\|_2 \) (\( \|c_n\|_2 \)) encourages the whole column (row) vector \( b_n \) (\( c_n \)) to be zero. The second sum in (31) penalizes individual entries of \( b_n \) and \( c_n \); thus, if \( b_n \neq 0 \) (\( c_n \neq 0 \)), then \( \|b_n\|_1 \) (\( \|c_n\|_1 \)) induces sparsity in its entries. Altogether, \( \Gamma(B, C) \) can set columns and rows of \( B \) and \( C \) to zero (depending on \( \gamma \)), and also individual entries of these columns and rows to zero (depending on \( \lambda \)).

### B. Majorizing the \( \Gamma \)-regularized stress

In this section, a convex and differentiable (with respect to \( X \)) majorizer \( g_t \) is constructed for \( f_t \). Thus, focus is placed on the first term of \( f_t \) while the last two terms remain unchanged. Although \( \Delta \) and \( D(X) \) are symmetric, \( B + C \) is not necessarily symmetric. Upon introducing the auxiliary \( N \times N \) matrix \( E := \Delta - (B + C) \), and expanding the squared norm, the first term of \( f_t \) in (30) can be compactly written as

\[
\frac{1}{2} \|E - D(X)\|_F^2 = \frac{1}{2} \|E\|_F^2 + \text{Tr}(XLX^T) - \frac{1}{2} \text{Tr}(E^T D(X) + D^T(X)E)
\]

where \( \text{Tr}(XLX^T) \) follows from \( \|D(X)\|_F^2/2 \) after straightforward algebraic manipulations [cf. (16)].

Let \( B_U \) and \( C_U \) (\( B_L \) and \( C_L \)) denote the upper (lower) triangular parts of \( B \) and \( C \), respectively. Define also the symmetric matrices \( B_U := B_U + B_U^T \), \( C_U := C_U + C_U^T \), \( B_L := B_L + B_L^T \), and \( C_L := C_L + C_L^T \). The following proposition (see proof in Appendix D) specifies a majorizer for the last term in (32).

**Proposition 5.** A majorizer for \( \frac{1}{2} \text{Tr}(E^T D(X) + D^T(X)E) \) is given by

\[
h_0(B, C, X; Z) = \frac{1}{2} g_0 \left( B_U + C_U, X; Z \right) + \frac{1}{2} g_0 \left( B_L + C_L, X; Z \right).
\]

The majorizer \( h_0 \) is defined through \( g_0 \) in Proposition 2 and accounts for both the upper and lower triangular parts of \( E \). Had \( B + C \) been symmetric, \( h_0 \) would reduce to \( g_0 \) in (19).

Finally, the majorizer \( g_t \) is obtained after substituting (32) and (33) into (30) to obtain

\[
g_t(B, C, X; Z) := \frac{1}{2} \|E\|_F^2 + \text{Tr}(XLX^T) + h_0(B, C, X; Z) + \Gamma(B, C) + \frac{\lambda}{2} \|B - C\|_F^2.
\]

Although \( g_t \) is not jointly convex with respect to \( (B, C, X) \), it is convex with respect to each one of \( B, C, \) and \( X \). This fact is instrumental to developing the solver presented in the ensuing section.

### C. Iterative solver based on generalized MM

The generalized MM updates to solve (29) are

\[
B^{(t+1)} = \arg \min_B g_t(B, C^{(t)}, X^{(t)}; X^{(t)})
\]
\[
C^{(t+1)} = \arg \min_C g_t(B^{(t+1)}, C, X^{(t)}; X^{(t)})
\]
\[
X^{(t+1)} = \arg \min_X g_t(B^{(t+1)}, C^{(t+1)}, X; X^{(t)})
\]

where \( B^{(t)} \) and \( C^{(t)} \) denote estimates of \( B \) and \( C \) at iteration \( t \), respectively. Since \( g_t \) majorizes \( f_t \), updates (35a) and (35b) reduce to

\[
B^{(t+1)} = \arg \min_B f_t(B, C^{(t)}, X^{(t)})
\]
\[
C^{(t+1)} = \arg \min_C f_t(B^{(t+1)}, C, X^{(t)}).
\]

Computing \( B^{(t+1)} \) via (36a) decouples across its columns \( \{b_n^{(t+1)}\}_{n=1}^N \). Let \( \kappa_n(t) \), \( \delta_n \), and \( d_n(X^{(t)}) \) denote the \( n \)-th column of \( C^{(t)} \), \( \Delta \), and \( D(X^{(t)}) \), respectively. Each column of \( B^{(t+1)} \) is obtained as \( b_n^{(t+1)} = \arg \min \phi_B(t) \) (cf. (37), (39) and Proposition 6), where

\[
\phi_B(t) := \frac{1}{2} \left\| \delta_n - d_n(X^{(t)}) - (\beta_n(t) + \kappa_n(t)) \right\|_2^2 + \gamma \|b_n\|_2 + \lambda \|b_n\|_1 + \mu \|b_n - c_n(t)\|_2^2.
\]

Although \( \phi_B(t) \) is convex, it is non-differentiable. Nevertheless, \( b_n^{(t+1)} \) can be obtained in closed form in terms of the vector soft-thresholding operator \( \mathcal{S}_{2\lambda}(b_n) := \left[ \mathcal{S}_{2\lambda}(b_{n,1}) \cdots \mathcal{S}_{2\lambda}(b_{n,N}) \right]^T \), as stated in the next proposition (see proof in Appendix E).

**Proposition 6.** The global minimizer of \( \min_{b_n} \phi_B(t)(b_n) \), denoted \( b_n^{(t+1)} \), is given in closed form by

\[
b_n^{(t+1)} = \mu_0^{-1} \left( 1 - \frac{\gamma}{\|\mathcal{S}_{2\lambda}(\hat{u}_n^{(t)})\|_2} \right) \mathcal{S}_{2\lambda}(\hat{u}_n^{(t)})
\]

where \( u_n^{(t)} := (\vec{\Delta} - d_n(X^{(t)}) - \kappa_n(t) + \mu \kappa_n(t)) \), and \( \mu_0 := 1 + \mu \).

Likewise, finding \( C^{(t+1)} \) via (36b) decouples across its rows \( \{c_n^{(t+1)}\}_{n=1}^N \). Let \( (\beta_n^{(t+1)})^T \) denote the \( n \)-th row of \( B^{(t+1)} \). Then, each row of \( C^{(t+1)} \) is found as \( c_n^{(t+1)} = \text{arg} \min_{c_n} \phi_C(t)(c_n) \), where

\[
\phi_C(t)(c_n) := \frac{1}{2} \left\| \delta_n - d_n(X^{(t)}) - (\beta_n^{(t+1)} + c_n) \right\|_2^2 + \gamma \|c_n\|_2 + \lambda \|c_n\|_1 + \mu \|b_n^{(t+1)} - c_n\|_2^2,
\]

Note that \( \delta_n \) and \( d_n(X^{(t)}) \) are common to (37) and (39) since \( \Delta \) and \( D(X^{(t)}) \) are symmetric. The global minimizer of (39) can be obtained in closed form as [cf. (37), (39) and Proposition 6]

\[
c_n^{(t+1)} = \mu_0^{-1} \left( 1 - \frac{\gamma}{\|\mathcal{S}_{2\lambda}(\hat{u}_n^{(t)})\|_2} \right) \mathcal{S}_{2\lambda}(\hat{u}_n^{(t)})
\]

where \( \hat{u}_n^{(t)} := \delta_n - d_n(X^{(t)}) - \beta_n^{(t+1)} + \mu b_n^{(t+1)} \).
Algorithm 2 Robust Structured MDS

Require: Set $B^{(0)}$ and $C^{(0)}$ to zero, and randomly initialize $X^{(0)}$.
1: for $t = 0, 1, 2, \ldots$ do
2:  Find each column of $B^{(t+1)}$ via (38).
3:  Find each row of $C^{(t+1)}$ via (40).
4:  Update $X^{(t+1)}$ via (43).
5: end for

Updates (38) and (40) illustrate the roles of $\gamma$ and $\lambda$ as sparsity controlling parameters. Note that if $\gamma > 0$ and $\lambda = 0$, then $\{b_{n}^{(t+1)}\}_{n=1}^{N}$ and $\{c_{n}^{(t+1)}\}_{n=1}^{N}$ with nonzero norm do not necessarily have a sparse structure. Likewise, if $\gamma = 0$ and $\lambda > 0$, then $B^{(t+1)}$ and $C^{(t+1)}$ have a sparse structure. However, their nonzero entries are not necessarily grouped along few rows and columns $B^{(t+1)}$ and $C^{(t+1)}$.

Updating $X^{(t+1)}$ via (35c) once again reduces to solving a quadratic program. The majorizer $g_{t}$ is now defined through matrices

$$
\tilde{L}_{1}(B^{(t+1)} + C^{(t+1)}, X^{(t)}) := L_{1}(B_{L}^{(t+1)} + C_{L}^{(t+1)}, X^{(t)}) + L_{1}(B_{U}^{(t+1)} + C_{U}^{(t+1)}, X^{(t)})
$$

Then, the first-order optimality condition for (35c) yields the closed-form solution

$$
X^{(t+1)} = X^{(t)} \tilde{L}_{1}(B^{(t+1)} + C^{(t+1)}, X^{(t)}) (\tilde{L}(t))^{†}
$$

where $\tilde{L}(t) := L + \tilde{L}_{2}(B^{(t+1)} + C^{(t+1)}, X^{(t)})$. The resulting embedding $X^{(t+1)}$ is forced to be centered at the origin, thereby removing the translational ambiguity with respect to $X$ present in $g_{t}$ (cf. (27)).

The robust structured MDS (RSMDS) scheme is summarized as Algorithm 2. It terminates when $\|X^{(t+1)} - X^{(t)}\|_F / \|X^{(t)}\|_F$ drops below a small positive threshold $\nu$. Let $(B, C, X)$ denote a limit point for $\{(B^{(t)}, C^{(t)}, X^{(t)})\}_{t \geq 1}$. Since $f_{t}$ is continuous, bounded from below, and $\{f_{t}(B^{(t)}, C^{(t)}, X^{(t)})\}_{t \geq 1}$ is monotonically decreasing, the following convergence result holds (cf. Proposition 4).

Proposition 7. The sequence $\{f(B^{(t)}, C^{(t)}, X^{(t)})\}_{t \geq 1}$ generated by RSMDS converges to $f(B, C, X)$.

The behavior of $\{(B^{(t)}, C^{(t)}, X^{(t)})\}_{t \geq 1}$ and its set of limit points $\Omega_{F}$ can be characterized following the steps used in Proposition 3 (not detailed here due to space limitations). Lastly, it is argued next that the computational complexity and memory requirement of RSMDS are similar to those of SMACOF.

Remark 5. (Time and space complexity) Per iteration RSMDS computes $D(X^{(t)})$, which incurs $O(pN^{2})$ complexity. Then, each column of $B$ is updated via (38) with $O(N^{2})$ complexity. Next, each row of $C$ is updated through (40) entailing $O(N^{2})$ computations as well. Then, $X$ is updated via (43) which involves computing two matrix multiplications and a matrix pseudoinverse with complexity $O(N^{3})$ in the worst case. Matrices $L_{1}$ and $L_{2}$ are also updated per iteration with $O(N^{2})$ operations. Hence, the total computational complexity per iteration is $O(N^{2})$. The space requirements of RSMDS are: (i) a $p \times N$ matrix $X$; (ii) two sparse $N \times N$ matrices, namely $B$ and $C$; and (iii) two $N \times N$ symmetric matrices $L$, $L_{1}$, and $L_{2}$, with $N(N+1)/2$ scalars each. Hence, RMDS requires $O(N^{2})$ memory.

V. NUMERICAL TESTS

Numerical performance of the robust MDS algorithms is tested in this section. First, RMDS is tested on synthetic data for an arbitrary distribution of outliers (Section V-A). Then, RMDS is used on packet-delay data to visualize the unknown topology of a computer network (Section V-B). Next, the performance of RMDS with missing data is illustrated on a corpus of face images (Section V-C). Finally, RSMDS is used to identify structured outlier patterns in a social network dataset (Section V-D), and to visualize path latency in real network data (Section V-E).
A. Synthetic data with noise

A fixed configuration $X$ of $N = 25$ points in a $p = 2$-dimensional space was considered for this test (see Fig. 2). The outlier-free data comprised $N_0 = 300$ noisy Euclidean distances $\delta_{n,m} = d_{n,m}(X) + \epsilon_{n,m}$, where $\{\epsilon_{n,m}\}$ were independent zero-mean random variables drawn from a zero-mean truncated Gaussian distribution with variance $\sigma^2 = 0.1$ and truncation threshold $\{-d_{n,m}(X)\}$ so that $\delta_{n,m} > 0$. The indexes $(n,m)$ of the outliers were chosen at random and their values were drawn independently from a uniform distribution over $[0,20]$. Several values of outlier contamination $S \in \{15,30,45,60,75\}$, corresponding to $\{5\%, 10\%, 15\%, 20\%, 25\%\}$ of the total distances, were considered.

RMDS was used to find a two-dimensional embedding based on the outlier contaminated data $\Delta$. Parameter $\lambda$ was chosen through three different criteria. First, assuming that $\sigma^2$ was known, the tuning parameter was set to $\lambda_1 := \lambda = 2.69\sigma$ (cf. Remark 1). Second, it was assumed that the degree of sparsity (number of outliers) $S$ was known. Then, RMDS was ran for all $\lambda \in \{\lambda_g\}_{g=1}^G$ as described in Remark 4. The parameter $\lambda_0 := \lambda$ was chosen such that the number of nonzero entries in $O_g$ was $2S$. The third method also chose $\lambda \in \{\lambda_g\}_{g=1}^G$ assuming that $\sigma^2$ was known. For each $\lambda_g$ the distances $\delta_{n,m}$ identified as outliers through $O_g$ were removed. The outlier-free distances were used to obtain an estimate $\hat{\sigma}^2$ as in (28). Then, $\lambda_{g^*} := \lambda_{g^*}$, with $g^* = \arg\min_g (\sigma^2 - \hat{\sigma}^2_g)^2$. The grid $\{\lambda_g\}_{g=1}^G$ comprised $G = 2,000$ points evenly spaced on a logarithmic scale. The upper limit of the grid’s range $\lambda_{\text{max}}$ was set to $\lambda_{\text{max}} = 2\max(n,m) [\delta_{n,m} - d_{n,m}(\hat{X}_{\text{SMACOF}})]$, where $\hat{X}_{\text{SMACOF}}$ denotes the embedding obtained by SMACOF [10]. The lower limit of the grid’s range was set to $\lambda_{\text{min}} = 10^{-4}\lambda_{\text{max}}$.

Both SMACOF and REE were used as benchmarks [23], [6]. YALMIP was used to find the REE embedding through the solution of its semidefinite program formulation [24]. It was empirically observed that the embedding produced by REE underestimated the true Euclidean distances as illustrated by the Shepard diagram in Fig. 1(a). One could correct the scaling of $X_{\text{REE}}$ by fitting a line through the REE scatter points in the Shepard diagram shown in Fig. 1(a). Then, the robust estimate $\hat{a}$ for the slope of the line, obtained via weighted LS with the Huber loss having parameter 1.345, was used to scale $X_{\text{REE}}$ and obtain the scaled REE (sREE) embedding $X_{\text{sREE}} := \hat{a}^{-1}X_{\text{REE}}$. Fig. 1(b) shows the Shepard diagram for sREE and compares it with the one obtained by RMDS. Although the pair-wise distances produced by the sREE embedding clustered around the line $\delta_{n,m} = d_{n,m}(X)$, they approximated the given $\{\delta_{n,m}\}$ with high variance; see also Fig 2. On the other hand, RMDS yielded an embedding whose $\{d_{n,m}(X)\}$ tightly clustered around the line $\delta_{n,m} = d_{n,m}(X)$.

The quality of the resulting embeddings was assessed through the normalized outlier-free stress

$$\sigma(X, \hat{O}) := \sqrt{\frac{\sum_{(n,m) \in T_0} (\delta_{n,m} - d_{n,m}(X))^2}{\sum_{(n,m) \in T_0} \delta_{n,m}^2}} \tag{44}$$

which measures the fit between $\{\delta_{n,m}\}$ and the pair-wise distances $\{d_{n,m}(X)\}$ obtained from the embedding only for distances not deemed as outliers. The minimum values of $\sigma(X, \hat{O})$ obtained for different levels of outlier contamination and $\lambda_0$ chosen as in Table I for each $S/N_0$.
### TABLE I

<table>
<thead>
<tr>
<th>$S$ ($S/N_0$)</th>
<th>$\lambda_0$</th>
<th>$\sigma(\hat{X}, \hat{O})$</th>
<th>$\hat{S}$</th>
<th>$\lambda_0$</th>
<th>$\sigma(\hat{X}, \hat{O})$</th>
<th>$\hat{S}$</th>
<th>$\sigma(\hat{X}^{\text{SMACOF}})$</th>
<th>$\sigma(\hat{X}^{\text{sREE}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 (5%)</td>
<td>0.851</td>
<td>0.0310</td>
<td>29</td>
<td>1.132</td>
<td>0.0350</td>
<td>15</td>
<td>4.8261</td>
<td>0.0590</td>
</tr>
<tr>
<td>30 (10%)</td>
<td>0.851</td>
<td>0.0304</td>
<td>38</td>
<td>0.8660</td>
<td>0.0328</td>
<td>30</td>
<td>2.3347</td>
<td>0.0787</td>
</tr>
<tr>
<td>45 (15%)</td>
<td>0.851</td>
<td>0.0315</td>
<td>62</td>
<td>1.1020</td>
<td>0.0394</td>
<td>45</td>
<td>2.2111</td>
<td>0.0589</td>
</tr>
<tr>
<td>60 (20%)</td>
<td>0.851</td>
<td>0.0327</td>
<td>81</td>
<td>1.0875</td>
<td>0.0442</td>
<td>60</td>
<td>1.7533</td>
<td>0.0581</td>
</tr>
<tr>
<td>75 (25%)</td>
<td>0.851</td>
<td>0.0314</td>
<td>93</td>
<td>1.0343</td>
<td>0.0374</td>
<td>75</td>
<td>1.8182</td>
<td>0.0574</td>
</tr>
</tbody>
</table>

In this section, a dataset of time delay differences between packets transmitted on a computer network of unknown topology was considered [7]. The goal was to construct a $p = 2$-dimensional map illustrating path overlaps between a fixed source-computer and $N = 10$ terminals in the network. A measurement scheme called ‘sandwich probing’, which sends three packets in a sequence, was used to obtain packet delays. Considering a fixed pair of terminal nodes $(n, m)$, a small packet is first sent to node $n$ followed by a large packet sent to node $m$. Finally, a small packet is sent again to node $n$. If the path from the source to both $n$ and $m$ overlaps, then the large packet undergoes delay proportional to the size of the overlap between the arrival of the two small packets to node $n$. Intuitively, the overlap of routing paths is inversely proportional to the distance (hops) between the corresponding terminals. The probes were send 9,567 times, including swaps between the large and small packet destinations for each pair of computers $(n, m)$ [7]. Fig. 3(a) depicts mean delays and their corresponding maximum and minimum values measured for all $N_0 = 45$ computer pairs.

The mean packet-delays $\tau_{n,m}$ between computer pairs were considered as the outlier-free data (see Fig. 3(a)), and its 2-dimensional embedding obtained via SMACOF is shown in Fig. 3(b). Note that $\tau_{n,m}$’s correspond to similarities between paths. Hence, they were transformed to dissimilarities $\delta_{n,m} = 100e^{-\tau_{n,m}/1,000}$. A similar transformation was applied to the maximum and minimum delays to obtain the smallest dissimilarity $\hat{\delta}_{n,m}^{\text{min}}$ and largest dissimilarity $\hat{\delta}_{n,m}^{\text{max}}$, respectively, for each pair of terminals. Data were artificially contaminated with outliers by picking outlier dissimilarity indexes $(n, m)$ uniformly at random. Then, outlier values were drawn from a uniform distribution over the interval $[\hat{\delta}_{n,m}^{\text{min}}, \hat{\delta}_{n,m}^{\text{max}}]$. Fig. 4(a) shows the normalized outlier-free stress value for RMDS and SMACOF when seeking embeddings of dimensionality $p = 1, 4, 7$ for the packet-delay data contaminated with $S \in \{2, 4, 6, 8, 10, 12\}$ outliers out of 45 dissimilarities. A total of 100 Monte Carlo runs with random initialization were performed for each $p$. In all cases, RMDS was able to identify all outliers correctly. Moreover, its performance was consistently better than SMACOF for all outlier contamination levels and all embedding-dimensions considered. Note that using $p > 4$ did not yield an embedding with improved fit, which hints towards using $p \leq 4$. The Shepard diagram for $p = 4$ and $S/N_0 = 12/45$ ($\approx 27\%$) comparing the embeddings obtained by RMDS, SMACOF and sREE (cf. Section V-A) is shown in Fig. 4(b). Again, RMDS produces an embedding whose pair-wise distances form a tighter cluster around the $\delta_{n,m} = d_{n,m}(X)$ line than SMACOF and sREE.
RMDS correctly identified all outliers and better approximated the non-outlier dissimilarities (see Fig. 3(b)).

Fig. 4(a) also illustrates the performance of RMDS when used for hard outlier rejection. In this case, the outliers identified by RMDS were removed from the dataset, and then SMACOF was run using the remaining (clean) dissimilarities. This method, termed RMDS-H, outperformed SMACOF, sREE, and RMDS. This is not surprising since RMDS identified the true outliers which were subsequently removed. The 2-dimensional embeddings obtained by SMACOF on the clean data (used as a benchmark), RMDS-H and sREE for the case of $S/N_0 = 2/45$ ($\approx 4.4\%$) outliers is shown in Fig. 3(b). Even visual inspection suggests that the embedding obtained by RMDS-H approximates better the SMACOF benchmark than sREE. Instead, one could consider the dissimilarities \( \{ \delta_{n,m}^s \}_{s=1}^S \) corresponding to the $S$ largest squared residuals \( \{ \delta_{n,m}^s - d_{n,m}(X_{\text{SMACOF}}) \}^2 \) as outliers. After comparing the \( \{ \delta_{n,m}^s \} \) with the true outliers, it turns out that the \( \{ \delta_{n,m} \} \) do not correspond to the true outliers. For instance, after fixing $p = 4$, the probability of correct outlier detection hovered around 50% for all levels of outlier contamination considered.

C. Face-pose images and missing data

In this section RMDS is used in a dimensionality reduction setup with missing entries in $\Delta$. The dataset comprised $N = 698$ face images of size $64 \times 64$ with different poses and illumination conditions [30]. It has been argued that albeit the images are in a high-dimensional space of dimensionality $4,096$, they are constrained to lie on a manifold of lower dimensionality [30]. To test this argument, it is prudent to obtain a two-dimensional embedding of the images while preserving their local structure summarized by pair-wise distances among them. The pair-wise dissimilarities (distances on the manifold) were approximated as follows. First, a complete weighted graph with $N = 698$ nodes was constructed. The edge weights were set equal to the Euclidean distance between the images at each end of the edge. For each node only the $4$ largest weighted-edges were kept yielding a graph that remained connected. Next, the geodesics between a subset of 99 node pairs were computed, as detailed in [30], to obtain some entries of $\Delta$. The remaining geodesics were considered missing. A fixed binary matrix $W \in \{0, 1\}^{N \times N}$ with entries $w_{n,m} := \|W\|_{n,m}$ being 1 if $\delta_{n,m}$ is available and zero otherwise, was introduced in the $\ell_1$-regularized stress to account for the missing entries (cf. Remark 2). A weighted version of RMDS using (9) was then implemented (cf. Section III-A).

The resulting $\Delta$ was artificially contaminated by $S = 6, 604$ and $S = 10, 098$ outliers which corresponds to roughly 10% and 15% of the available distances. The indexes $(n, m)$ for outlier dissimilarities $\delta_{n,m}$ were chosen uniformly at random over the set of indexes $\{(n, m) : w_{n,m} = 1\}$, and kept fixed for the experiment. Their value was drawn from a uniform distribution over the interval $[0, 3\delta_{n,m}^{\max}]$, where $\delta_{n,m}^{\max}$ was the largest geodesic available. Figs. 5(c) and 5(d) show the resulting embeddings obtained by RMDS for $S = 6, 604$ and $S = 10, 098$, respectively. As illustrated by the sample images shown on the embedding, RMDS preserves the embedding structure in the presence of outliers when compared with the benchmark embedding obtained by SMACOF on the clean dataset. On the other hand, using SMACOF on the outlier contaminated dataset yielded an embedding whose underlying structure was considerably affected by the outliers even in the case $S = 6, 604$ (see Fig. 5(b)). In the latter case, RMDS yielded $\sigma(X, O) = 0.151$ and SMACOF yielded $\sigma(X_{\text{SMACOF}}) = 0.372$.

D. Lloyds Bank Employees' Career data

In this section, RSMDs was used to obtain a two-dimensional map of the careers of $N = 80$ Lloyds Bank employees in the cohort of 1905-1909 [29]. Individual careers were seen as pair-sequences sampled yearly, where each pair comprised job location (branch type) and job-title held. Optimal matching was used to obtain dissimilarities between career paths of different employees. The resulting $\delta_{n,m}$ quantified the difficulty to transform one career path to another [1]. Allowed transformations included deletions, insertions and substitutions, each of them incurring a prescribed cost [29, Appendix A]. The $\delta_{n,m}$ between two sequences corresponded to the
minimum cost across all possible transformations standardized by the length of the longest career sequence. Note that some \( \delta_{n,m} = 0 \) for \( n \neq m \), and thus their corresponding points in the embedding overlap. Fig. 6(a) shows the embedding obtained by SMACOF on both the clean and the outlier contaminated data. Each point indicates the careerpath of an individual. The closer two points appear in the embedding, the more similar the careers paths of the corresponding individuals were. The dataset was artificially contaminated by introducing random outlier dissimilarities at random for fixed individuals \( n \in \{1, 9786\} \). Outliers were randomly drawn from a uniform distribution on \([0, 3\delta_{n,m}]\).

RSMDS requires positive parameters \( \gamma, \lambda, \) and \( \mu \). In this experiment, values for \( \mu \in \{1, 2, \ldots, 10\} \) were considered. Note that the Laplacian \( \hat{L}_2 \) in (43) was always zero for all \( \mu \) values considered. Hence, \( \hat{L}^T \) in (43) was computed only once at the beginning of the algorithm. Although \( \mu \) did not affect the final embedding obtained, it did affect the convergence speed of RSMDS: larger \( \mu \) values decreased the convergence speed. Also, \( \mu \) affected how close the final estimates \( \hat{B} \) and \( \hat{C}^T \) were to each other. Larger \( \mu \) yielded smaller \( \|\hat{B} - \hat{C}^T\|_F \). Once \( \mu \) was fixed, \( \gamma \) and \( \lambda \) were chosen through a two-dimensional grid search. Consider the one-dimensional grids \( \mathcal{L}_1 := \{\gamma_g\}_{g=1}^{G_1} \) and \( \mathcal{L}_2 := \{\lambda_g\}_{g=1}^{G_2} \) for the parameters \( \gamma \) and \( \lambda \), respectively. Parameters \( \lambda_{\max} \) and \( \lambda_{\min} \) which define the range of grid values for \( \mathcal{L}_2 \) were set to \( \lambda_{\max} := \max_{n,m} |\hat{d}_{n,m} - d_{n,m}(X_{\text{SMACOF}})| \) and \( \lambda_{\min} := 10^{-4}\lambda_{\max} \), respectively. Parameters \( \gamma_{\max} \) and \( \gamma_{\min} \) which define \( \mathcal{L}_1 \) were set based on grid points in \( \mathcal{L}_2 \) to \( \gamma_{\max}(\lambda_g) := \max_{n \in \{1, \ldots, N\}} \|\hat{S}_2\lambda_g(\hat{d}_{n} - d_{n}(X_{\text{SMACOF}}))\|_2 \) and \( \gamma_{\min} := 10^{-3}\gamma_{\max} \), respectively. Note that \( \hat{B} \) and \( \hat{C} \) are zero for any pair of tuning parameters \( (\gamma, \lambda) \) with either \( \gamma > 0 \) and \( \lambda \geq \lambda_{\max} \), or, \( \lambda > 0 \) and \( \gamma \geq \gamma_{\max}(\lambda) \). Then, RSMDS was run for all pairs \((\gamma_g, \lambda_g) \in \mathcal{L}_1 \times \mathcal{L}_2 \) using \( G_1 = 50 \) and \( G_2 = 50 \). The grid was traversed in a lexicographic order using warm starts to initialize each instance of RSMDS. Finally, prior knowledge of \( S \) and the number of objects introducing outliers was used to choose \((\gamma, \lambda)\). After setting \( \mu = 4 \), parameters \( \gamma = 8.9361 \) and \( \lambda = 0.9786 \) yielded \( S = 140 \), and the known number of objects introducing outliers.

RMSD with its tuning parameter chosen to identify \( S = 140 \) was used as a benchmark. Fig. 6(a) depicts the embedding obtained by both RSMDS and RMSD. Both algorithms yielded embeddings which preserved the structure of objects whose dissimilarities were not corrupted by outliers as illustrated by the indexed objects (landmarks). The normalized outlier-free stress values obtained were 0.2490 (SMACOF), 0.1713 (RMSD), and 0.1815 (RSMDS). Fig. 7(a) shows the outlier structure identified by RSMDS through the matrix \( \hat{B} + \hat{C} \). RSMDS was able to identify the outlier’s group structure.
while maintaining a performance comparable to that of RMDS. Note that the outlier matrix $\hat{O}$ obtained by RMDS did not capture the group structure of the outliers (see Fig. 7(b)). The corresponding Shepard diagram in Fig. 6(b) shows that RSMDS, unlike RMDS, considers whether the objects whose $\delta_{n,m}$ are under consideration were deemed as outliers. Hence, dissimilarities lying close to the line $\delta_{n,m} = d_{n,m}(X)$ were sentenced as outliers in spite of the presence of non-outlier $\{\delta_{n,m}\}$ located farther away from the same line.

E. Active Measurement Project networking data

In this section, RMDS and RSMDS were used to obtain a two-dimensional embedding of source-destination (SD) pairs based on latency data from the active measurement project (AMP) of the National Laboratory of Applied Network Research [25]. The AMP data comprised latency measurements for Internet Protocol version 4 (IPv4) packets on $N = 186$ SD pairs. Data was collected every 10 minutes between July 17, 2011 and August 17, 2011 yielding a $4,464 \times 1$ vector of latency measurements per SD path. The Euclidean distances between the latency measurement vectors per SD path were used to construct $\Delta$. Outliers occur in this data set due to dropped packets and communication errors, but also owing to the markedly different behavior of some SD paths.

To construct a two-dimensional embedding for the SD paths using RMDS and RSMDS, the tuning parameters $\lambda$, $\gamma$ and $\mu$ must be chosen. The challenge here is that an estimate of either the number of outliers present in the data or the noise variance is not available (cf. Remark 4). Instead, one can rely on geometric characteristics of the residuals induced by the presence of outliers. Fig. 8(a) shows the number of outliers identified by RMDS as a function of $\lambda$. This curve was obtained at a low computational cost by defining a grid of $\lambda$ values and using warm starts as described in Remark 4. At $\lambda^* = 355.5$, the curve showed an elbow which characterizes the transition from a region where outliers were identified to a region where nominal dissimilarities were deemed as outliers.

Fig. 8(b) shows the Shepard diagram obtained by RMDS and SMACOF, and validates this choice of $\lambda^*$ based on Fig. 8(a).

One could also use RSMDS, by considering the aggregate magnitude of outlier variables per object via $\zeta_m := \sum_{n=1}^N |\hat{o}_{n,m}|$. Computing $\{\zeta_m\}_{m=1}^{186}$ showed that most outliers corresponded to dissimilarities related to 8 objects as shown by the red points at the bottom of Fig. 9(a). Then, RSMDS parameters $\lambda^*$ and $\gamma^*$ were chosen via a two-dimensional grid search to yield 308 outliers and 8 objects responsible for the
outliers (cf. Section V-D). For this experiment, \( \mu = 4 \) was found to work well. Although several values for \( \mu \) were used, no noticeable changes were observed in the quality of the embedding or the number of outliers identified by RSMDS. The Shepard diagram obtained for RSMDS is shown in Fig. 9(b), and the resultant embedding in Fig. 10(a), along with sample SD latency sequences depicted in Fig. 10(b). The embedding reveals two groups of SD’s according to their latency patterns. The first group (see landmarks SD99, SD128, SD134) shows periodic traffic patterns where the largest latency periods relate to business hours between Monday and Friday. The second group (see landmarks SD50, SD170, SD183) shows stable latency patterns with few high latency bursts. Finally, the outliers depict volatile SD pairs with increased latency towards the tail of the records.

VI. CONCLUSIONS

Novel algorithms for robust MDS were developed based on a dissimilarity model explicitly accounting for outliers. Translating the unusual presence of outliers in the dissimilarities to sparsity in the domain of outliers led to a regularized LS problem formulation. The regularization term was chosen to control sparsity on the outlier variables, thereby establishing a link to the area of compressive sampling.

To develop tractable algorithms, the regularization term was surrogated by its closest convex envelope. Then, an iterative solver was developed via MM. The resulting updates involved scalar Lasso and grouped Lasso problems whose solutions per iteration become available in closed form. Moreover, the aggregate computational complexity of the novel algorithms was found of the same order as that of SMACOF, the popular non-robust MDS alternative. Numerical tests on artificially contaminated and real datasets illustrated the merits of the proposed algorithms.

The robust MDS formulation developed here can be extended to incorporate parametric transformation of the dissimilarities. This will lead to robustified versions of ratio, interval, logarithmic and exponential MDS. Moreover, the novel dissimilarity model can also be used with other stress functions, such as S-stress, or weighted versions of stress functions, such as Sammon mapping and elastic scaling. Although this work has focused on Euclidean distances, the family of Minkowsky distances is also of interest in areas such as psychology and sociology. The proposed algorithms can be modified to incorporate Minkowsky distances after redefining the majorizers \( g \) and \( g_{1/2} \) along the lines of [15]. Finally, as
shown in Section V-C, the robust MDS framework is readily applicable to Isomap whereby one can obtain robust Euclidean embeddings of data lying on manifolds embedded in high-dimensional spaces.

APPENDIX A
PROOF OF PROPOSITION 2

To construct a majorizer of \( \phi(O, X) := -2 \sum_{n < m} (\delta_{n,m} - o_{n,m})d_{n,m}(X) \), consider splitting \( \phi(O, X) \) into two sums depending on the sign of \( \delta_{n,m} - o_{n,m} \). Upon defining the auxiliary index sets \( Q_+ := \{(n, m) : n < m, \delta_{n,m} > o_{n,m}\} \) and \( Q_- := \{(n, m) : n < m, \delta_{n,m} \leq o_{n,m}\} \), it is possible to write \( \phi(O, X) = \phi_+(O, X) + \phi_-(O, X) \), where

\[
\phi_+(O, X) := -2 \sum_{(n, m) \in Q_+} (\delta_{n,m} - o_{n,m})d_{n,m}(X) \tag{45a}
\]
\[
\phi_-(O, X) := -2 \sum_{(n, m) \in Q_-} (\delta_{n,m} - o_{n,m})d_{n,m}(X). \tag{45b}
\]

A majorizer of \( \phi_+ \) is obtained by using the Cauchy-Schwartz inequality \((z_m - z_n)^T(x_n - x_m) \leq d_{n,m}(Z)d_{n,m}(X) \) on each summand of (45a) to verify that

\[
\phi_+(O, X) \leq -2 \sum_{(n, m) \in Q_+} \frac{\delta_{n,m} - o_{n,m}}{d_{n,m}(Z)}(z_n - z_m)^T(x_n - x_m) \leq -2 \sum_{(n, m) \in Q_+} \delta_{n,m}d_{n,m}(Z)T(x_n - x_m) \tag{46}
\]

where equality holds whenever \( Z = X \). Note that if \( d_{n,m}(Z) = 0 \) for some \( n, m \in \{1, \ldots, N\} \), then the r.h.s. of (46) becomes undefined. In this case, the trivial bound \( 0 \geq d_{n,m}(X) \) is used.

Although the summands in \( \phi_- \) are convex with respect to \( X \), they remain non-differentiable. A majorizer for \( \phi_- \) is obtained by using the fact that \( \sqrt{X} \) is upper bounded as \( \sqrt{X} \leq \sqrt{X_0} + 0.5x_0^{-1/2}(x - x_0) \) at any \( x > 0 \). Equality holds in the latter if \( x = x_0 \). Letting \( x = d_m^2(X) \) and \( x_0 = d_m^2(Z) \), and using the bound on each summand of (45b) leads to

\[
\phi_-(O, X) \leq -2 \sum_{(n, m) \in Q_-} \frac{\delta_{n,m} - o_{n,m}}{d_{n,m}(Z)}d_{n,m}(X) + C_0(O, Z) \tag{47}
\]

where

\[
C_0(O, Z) := -2 \sum_{(n, m) \in Q_-} \delta_{n,m}d_{n,m}(Z). \tag{48}
\]

Again, if \( d_{n,m}(Z) = 0 \) in (47) for some \( n, m \in \{1, \ldots, N\} \), then the trivial bound \( 0 \geq d_{n,m}(X) \) is used instead.

A majorizer \( g_0(O, X; Z) \geq \phi(O, X) \) is obtained after adding (46) to (47). Using (18) and after some algebraic manipulations, a compact form for \( g_0 \) follows

\[
g_0(O, X; Z) = -2\text{Tr}(XL_1(O, Z)Z^T) + \text{Tr}(XL_2(O, Z)X^T) + C_0(O, Z) \tag{49}
\]

thereby concluding the proof.

APPENDIX B
PROOF OF PROPOSITION 3

The proof begins by establishing the following auxiliary result, which follows from [18].

**Lemma 2.** If \((O^*, X^*)\) is a local minimum of \( f \) in (7), then \( d_{n,m}(X^*) = 0 \) if and only if \( d_{n,m} = o_{n,m}^* \).

The usable local minima \((O^*, X^*)\) of \( f \) in (7) can be characterized through

\[
0_{N \times N} \in \partial_0 f(O^*, X^*) \tag{50a}
\]
\[
X^*L = X^*L_1(O^*, X^*) \tag{50b}
\]

where \( \partial_0 f(O^*, X^*) \) denotes the subdifferential of \( f \) with respect to \( O \) evaluated at \((O^*, X^*)\), and (50b) follows since Lemma 2 implies that \( f \) is differentiable with respect to \( X \) at each \((O^*, X^*)\).

Let \((\bar{O}, X)\) denote a fixed point of (22). Since (22a) is equivalent to (23), at a fixed point (22a) satisfies \( 0_{N \times N} \in \partial_0 f(O, X) \). Also, at a fixed point (22b) satisfies \( XL = XL_1(O, X) \) (cf. (27)). By Lemma 1, all fixed points of (22) are usable. Hence, all \((\bar{O}, X)\) satisfy (50a) and (50b), and are thus usable local minima for (7) as established in the following lemma.
Lemma 3. The fixed points of (22) are usable local minima of (7).

The convergence of the sequence \( \{(\mathbf{O}^{(i)}, \mathbf{X}^{(i)})\}_{i \geq 1} \) generated by (22) is examined next. Let \( \Omega := S_N^N \times \mathbb{R}^{p \times N} \) denote the domain of \( g \), and define the point-to-set maps \( M_1(\mathbf{O}^{(t)}, \mathbf{X}^{(t)}) := \{ (\mathbf{O}, \mathbf{X}) \in \Omega \} \) and \( M_2(\mathbf{O}^{(t)}, \mathbf{X}^{(t)}) := \{ (\mathbf{O}, \mathbf{X}) \in \Omega : \mathbf{X} \mathbf{1}_{1:N} = \mathbf{0}_{N} \} \). Also, let \( \omega := \{ (\mathbf{O}, \mathbf{X}) \} \), \( \omega(\tau) := (\mathbf{O}^{(\tau/2+1)}, \mathbf{X}^{(\tau/2)}) \) if \( \tau \) is odd, and \( \omega(\tau) := (\mathbf{O}^{(\tau/2+1)}, \mathbf{X}^{(\tau/2+1)}) \) if \( \tau \) is even; and

\[
g(\omega(\tau)) := \begin{cases} g(\mathbf{O}^{(\tau/2+1)}, \mathbf{X}^{(\tau/2+1)}), & \text{if } \tau \text{ odd;} \\
g(\mathbf{O}^{(\tau/2+1)}, \mathbf{X}^{(\tau/2)}), & \text{if } \tau \text{ even.} \end{cases} \tag{51}
\]

For a given initialization \( \omega(0) \) and since the sequence \( \{g(\omega(\tau))\}_{\tau \geq 1} \) generated by (22) is monotonically decreasing, then all \( \omega(\tau) \) belong to \( H_0 := \{ \omega \in \Omega : f(\mathbf{O}, \mathbf{X}) \leq g(\omega(0)) \} \). Since \( f \) is continuous, its epigraph is closed; hence, all its level sets are closed too. In particular, \( H_0 \) is closed. Moreover, for a given \( \omega(0) \) the set \( H_0 \) is bounded and, thus, compact. Hence, the following result follows by the Bolzano-Weierstrass Theorem [22, Ch. 2].

**Lemma 4.** The sequence \( \{\omega(\tau)\}_{\tau \geq 1} \) has at least one limit point in \( \Omega \).

Each update in (22) can be written in compact form in terms of the operators \( \mathbf{U}_1 : H_0 \rightarrow \{ \omega(\tau+1) \in M_1(\omega(\tau)) : g(\omega(\tau+1)) \leq g(\omega(\tau)) \} \) and \( \mathbf{U}_2 : H_0 \rightarrow \{ \omega(\tau+1) \in M_2(\omega(\tau)) : g(\omega(\tau+1)) \leq g(\omega(\tau)) \} \) which are defined through the closed-form updates (25) and (27), respectively. Given a fixed initialization \( \omega(0) \), the iterations (22) can be compactly written as \( \omega(\tau+1) = \mathbf{U}_1(\omega(\tau)) \), for \( \tau = 0, 1, \ldots \), where \( \mathbf{v}(\tau) := (\tau \mod 2) + 1 \) is a variable selection operator. The following lemma shows that the sequence generated by \( \omega(\tau+1) = \mathbf{U}_1(\omega(\tau)) \) is asymptotically regular, which means that \( \|\omega(\tau+1) - \omega(\tau)\| \rightarrow 0 \) as \( \tau \rightarrow \infty \), and that its limit points are fixed points of (22).

**Lemma 5.** The sequence \( \{\omega(\tau)\}_{\tau \geq 1} \) is asymptotically regular. Moreover, every limit point \( \bar{\omega} \in \{\omega(\tau)\}_{\tau \geq 1} \) satisfies \( \bar{\omega} = \mathbf{U}_1(\bar{\omega}) = \mathbf{U}_2(\bar{\omega}) \); hence, there is a fixed point of (22).

**Proof:** First it is shown that \( g \) and \( \{M_i\}_{i=1}^2 \) satisfy the following properties: \( (P1) \) \( \omega \in M_i(\omega) \), \( \forall \omega \in \Omega, i = 1, 2 \); \( (P2) \) \( M_i(\omega) \) is continuous on \( \Omega, i = 1, 2 \); and \( (P3) \) \( \forall \omega \in \Omega, g \) has a unique minimizer over each \( M_i(\omega), i = 1, 2 \). \( (P4) \) \( \exists \omega(0) \) \( \in \Omega \) such that \( H_0 \) is a compact subset; and, \( (P5) \) \( \exists \tau, \tau' \geq 2 \) such that \( 0 < \tau < \tau' \) and \( \exists \tau \in [1, \tau'] \) so that \( \omega(\tau+1) = \mathbf{U}_1(\omega(\tau)) \). Property (P1) is automatically satisfied by the definition of \( M_1 \) and \( M_2 \). Property (P2) follows since both \( M_1 \) and \( M_2 \) are continuous mappings. The minimizer of \( g \) with respect to \( \mathbf{O} \) over \( M_1 \) corresponds to the solution of a batch of scalar Lasso problems; thus, the minimizer is unique. The minimizer of \( g \) with respect to \( \mathbf{X} \) over \( M_2 \) corresponds to the minimum-norm solution for \( \mathbf{X} \) and it is, thus, unique. Hence, Condition (P3) holds. Property (P4) is automatically satisfied. Properties (P1)-(P4) and [11, Property 11] guarantee that \( \{\omega(\tau)\}_{\tau \geq 1} \) is asymptotically regular. The cyclic update rule defined by \( \mathbf{v}(\tau) \) is a special case of (P5) with \( \tau = 2 \). Hence, the second part of the result follows by [11, Thm. 15].
across entries of $w$, and each entry of $\hat{w}$ is given in closed form as
\[
[\hat{w}]_m = \begin{cases} 
\frac{u(t)}{(n,m)} \lambda & |u(t)|_{(n,m)} \leq \lambda \\
\sgn(u(t))_{(n,m)} & |u(t)|_{(n,m)} > \lambda 
\end{cases}.
\] (56)

Next, if $1 \geq \|\hat{v}\|_2 = \|u(t) - \lambda \hat{w}\|_2/\gamma$ holds true, which is equivalent to
\[
\|u(t) - \lambda \hat{w}\|_2 \leq \gamma
\] (57)
then $b^{(t+1)}_n = 0_N$. Otherwise, $b^{(t+1)}_n \neq 0_N$ and the subgradient $v$ is unique.

For the case $b^{(t+1)}_n \neq 0_N$, let $\mu_0 := 1 + \mu$, and set $v = b^{(t+1)}_n/\|b^{(t+1)}_n\|_2$ in (54). The cost $g^{(t)}_{b^{(t+1)}_n}$ remains non-differentiable due to $|b^{(t+1)}_n|_1$. Considering each entry of $b^{(t+1)}_n$ separately leads to the following two cases. First, if $[b^{(t+1)}_n]_m \neq 0$, the corresponding subgradient entry in (54) yields
\[
[b^{(t+1)}_n]_m \left(\mu_0 + \frac{\gamma}{\|b^{(t+1)}_n\|_2}\right) = u(t)_{(n,m)} - \sgn(u(t))_{(n,m)} \lambda.
\] (58)

Else, if $[b^{(t+1)}_n]_m = 0$, then (54) leads to $-u(t)_{(n,m)} + \lambda |w|_m = 0$, where $|w|_m \leq 1$, which yields
\[
|u(t)_{(n,m)}| \leq \lambda.
\] (59)

Considering the vector form of (58) and (59) and using the vector soft-thresholding operator yields
\[
b^{(t+1)}_t \left(\mu_0 + \frac{\gamma}{\|b^{(t+1)}_t\|_2}\right) = S_{2\lambda} \left(u(t)\right).
\] (60)

Equation (60) reveals that $b^{(t+1)}_n$ is a scaled version of $S_{2\lambda} \left(u(t)\right)$. The scaling can be found by taking the $\ell_2$-norm on both sides of (60) from which one obtains $\|b^{(t+1)}_n\|_2 = (\|S_{2\lambda} \left(u(t)\right)\|_2 - \gamma)/\mu_0$. Replacing $[b^{(t+1)}_n]_m$ in (60), solving for $b^{(t+1)}_n$, and introducing (57) through the $\lambda$ notation leads to (38), and concludes the proof.

References


Pedro A. Forero (S’07) received his Diploma in Electronics Engineering from Pontificia Universidad Javeriana, Bogota, Colombia in 2003 and M.Sc. degree in Electrical Engineering from Loyola Marymount University in 2006. Since September 2006, he has been working towards his Ph.D. degree with the Department of Electrical and Computer Engineering at the University of Minnesota. His research interests include statistical signal processing, machine learning and network science. His current research focuses on robust algorithms for manifold learning and traffic prediction over complex networks.

Mr. Forero has been a recipient of the Science, Mathematics and Research for Transformation (SMART) scholarship since 2007.

G. B. Giannakis (IEEE Fellow’97) received his Diploma in Electrical Engr. from the Ntl. Tech. Univ. of Athens, Greece, 1981. From 1982 to 1986 he was with the Univ. of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engr., 1986. Since 1999 he has been a professor with the Univ. of Minnesota, where he now holds an ADC Chair in Wireless Telecommunications in the ECE Department, and serves as director of the Digital Technology Center.

His general interests span the areas of communications, networking and statistical signal processing - subjects on which he has published two edited books, two research monographs, 20 book chapters, 325 journal and 550 conference papers (H-index=95). Current research focuses on compressive sampling, cognitive radios, cross-layer designs, wireless sensors, social and power grid networks. He is the (co-) inventor of 21 patents issued, and the (co-) recipient of 8 best paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received Technical Achievement Awards from the SP Society (2000), from EURASIP (2005), a Young Faculty Teaching Award, and the G. W. Taylor Award for Distinguished Research from the University of Minnesota. He is a Fellow of EURASIP, and has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SP Society.