ROBUST MULTIDIMENSIONAL SCALING EMPLOYING M-ESTIMATORS AND NUCLEAR NORM REGULARIZATION

Fotios Mandanas and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki, 54124, Greece
Email: {fmandan@gmail.com, costas@aiia.csd.auth.gr}

ABSTRACT

Multidimensional Scaling (MDS) is applied to pairwise dissimilarities between entities, aiming to map each entity to a point in a geometric space so that the inter-point distances preserve the pairwise dissimilarities. The well-known algorithms for solving the MDS problem are vulnerable to gross errors (outliers), inducing highly corrupted embeddings. To cope with such gross errors, two algorithms are proposed, which resort to half-quadratic optimization, employing M-estimators and nuclear norm regularization. It is demonstrated by experiments that the proposed algorithms outperform the state-of-the-art MDS ones.

***Index Terms*—** Multidimensional scaling, robustness, M-estimators, nuclear norm, half-quadratic optimization

1. INTRODUCTION

Multidimensional Scaling (MDS) offers a visualization of the hidden structures among a set of entities in a geometric space of reduced dimensions, preserving the pairwise dissimilarities between entities. Its input is a square symmetric dissimilarity matrix that captures the dissimilarities among the set of entities, while its output is a model in a geometric space of two or three dimensions, where each entity is represented by a single point. A spectrum of MDS applications can be found in [1]. Common techniques solving the MDS problem, such as the classical MDS [2] and the scaling by majorizing a complicated function (SMACOF) [3], have shown to be less robust, when the pairwise dissimilarities are corrupted by gross errors [1,4].

Here, we advocate that the exploitation of M-estimators in the MDS algorithm with a proper regularization can mitigate the repercussion of gross-errors more efficiently than the state-of-the-art techniques. The paper contributions are: 1) The extension of the framework proposed in [1], that is based on half-quadratic (HQ) optimization, with two algorithms, which employ M-estimators and nuclear norm to impose smoothness whenever the initial dissimilarity matrix is contaminated by gross errors. 2) The demonstration of the benefits of the proposed algorithms against the sophisticated MDS techniques.

The underlying reasoning for the use of nuclear norm (also known as trace norm, Schatten 1-norm, or Ky Fan r-norm) stems from the rich related literature. The nuclear norm is the best convex approximation of the rank function over the unit ball of matrices with norm less than one [5]. A singular value thresholding algorithm for matrix completion and related nuclear norm minimization problems is proposed in [6]. In [5], the NP-hard affine rank minimization problem is solved. Especially, if a specific restricted isometry property holds for the linear transformation that defines the constraints, the minimum rank solution problem turns to be the solution of the convex nuclear norm minimization problem. In [7], the data matrix is assumed to be the superposition of a low-rank and a sparse component. It is proven that, under certain assumptions, a disentanglement of both low-rank and sparse components is achieved by solving a convex program called Principal Component Pursuit. However, most nuclear norm minimization techniques exploit singular value thresholding algorithms at each iteration, which induces additional computational cost as the matrix size increases. Thus, a Schatten p-norm optimization framework for the solution of rank and trace norm objectives is proposed in [8], yielding a closed-form solution suitable for large-scale matrix completion problems.

For $\mathbf{x} \in \mathbb{R}^{d\times 1}$, $\|\mathbf{x}\|_1 = \sum_{i=1}^{d}|x_i|$ and $\|\mathbf{x}\|_2 = \sqrt{\sum_i^d x_i^2}$ are the ℓ_1 and ℓ_2 norms of $\mathbf{x}$, respectively. Let $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|,\ldots,|\mathbf{x}_N]^T \in \mathbb{R}^{N\times d}$, where the i-th row of $\mathbf{X}$ is denoted as $\mathbf{x}^i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{id}) \in \mathbb{R}^{1\times d}$. The Frobenius norm of $\mathbf{X}$ is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{N}\|\mathbf{x}^i\|_2^2}$. For $N > d$, the nuclear norm of $\mathbf{X}$ is defined as the sum of its singular values, namely $\|\mathbf{X}\|_* = \mathrm{tr}((\mathbf{X}^T\mathbf{X})^{1/2}) = \sum_{i=1}^{d}\sigma_i$, constituting a special case of the Schatten norm $\|\mathbf{X}\|_p = (\sum_{i=1}^{d}\sigma_i^p)^{\frac{1}{p}}$. The subdifferential of the nuclear norm at $\mathbf{X}$, $\partial\|\mathbf{X}\|_*$, is $\mathbf{U}\mathbf{V}^T$, where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the singular value decomposition (SVD) of $\mathbf{X}$ [9].

2. ROBUST MDS APPROACHES

Let N denote the number of entities, d be the reduced embedding dimension, and $\mathbf{\Delta} = [\delta_{ij}]$ denote the pairwise dissimilarity matrix with δ_{ij}, $i,j = 1,2,\ldots,N$ corresponding to the dissimilarity between the entities i and j. The embedding in the reduced d dimensional space is declared as $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|,...,|\mathbf{x}_N]^T \in \mathbb{R}^{N\times d}$ with the i-th object mapped to $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{id})^T \in \mathbb{R}^{d\times 1}$. The distance matrix is denoted as $\mathbf{D}(\mathbf{X}) = [d_{ij}(\mathbf{X})] \in \mathbb{R}^{N\times N}$ with ij-th element being equal to ℓ_2 norm between $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e., $d_{ij}(\mathbf{X}) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$.

MDS objective is to estimate $\mathbf{X}$ by minimizing the least-squares (LS) loss function of raw stress, which is vulnerable to gross errors:

$$\sigma_r(\mathbf{X}) = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(\delta_{ij} - d_{ij}(\mathbf{X}))^2 \triangleq \sum_{i<j}^{N}(\delta_{ij} - d_{ij}(\mathbf{X}))^2. \quad (1)$$

Under the context of minimizing outliers impact, the function $\left\|\mathbf{\Delta}^2 - \mathbf{D}^2\right\|_1$ was employed in the robust Euclidean embedding (REE) [10]. Additional schemes can be found in [11, 12]. In the robust MDS (RMDS) [4], each dissimilarity element is modeled as $\delta_{ij} = d_{ij}(\mathbf{X}) + o_{ij} + \epsilon_{ij}$, where o_{ij} represents a gross error, while

ϵ_{ij} denotes a zero-mean independent random variable modeling the nominal errors. Since the gross errors are sparse, the ℓ_1 norm is enforced to them, yielding [4]:

$$(\hat{\mathbf{O}}, \hat{\mathbf{X}}) = \underset{\mathbf{O},\mathbf{X}}{\operatorname{argmin}} \sum_{i<j}^{N} (\delta_{ij} - d_{ij}(\mathbf{X}) - o_{ij})^2 + \lambda_1 \sum_{i<j}^{N} |o_{ij}|. \quad (2)$$

An iterative solution for (2) is:

$$o_{ij}^{(t+1)} = S_{\lambda_1}(\delta_{ij} - d_{ij}(\mathbf{X}^{(t)})) \quad (3)$$

$$\mathbf{X}^{(t+1)} = \mathbf{L}^{\dagger}\, \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\, \mathbf{X}^{(t)} \quad (4)$$

where $S_{\lambda_1}(x) = \operatorname{sign}(x)(|x| - \frac{\lambda_1}{2})_+$ is the soft-thresholding operator with $(\cdot)_+ = \max\{\cdot, 0\}$. Since the matrix $\mathbf{L}$, with diagonal elements $[\mathbf{L}]_{ii} = N-1$ and off-diagonal elements $[\mathbf{L}]_{ij} = -1$, is not full rank, its Moore-Penrose pseudoinverse $\mathbf{L}^{\dagger} = N^{-1}\mathbf{J}$ is used, where $\mathbf{J} = \mathbf{I} - N^{-1}\,\mathbf{e}\,\mathbf{e}^T$ is the centering operator and $\mathbf{e}$ is the $N \times 1$ vector of ones. The Laplacian matrix $\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})$ is defined:

$$[\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})]_{ij} = \begin{cases} -(\delta_{ij} - o_{ij})\, d_{ij}^{-1}(\mathbf{X}) & (i,j) \in \mathbb{S}(\mathbf{O}, \mathbf{X}) \\ 0 & (i,j) \in \mathbb{T}(\mathbf{O}, \mathbf{X}) \\ -\sum_{k=1, k\neq i}^{N} [\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})]_{ik} & (i,j) \in \mathbb{Q}(\mathbf{O}, \mathbf{X}) \end{cases} \quad (5)$$

where $\mathbb{S}(\mathbf{O}, \mathbf{X}) = \{(i,j) : i \neq j,\, d_{ij}(\mathbf{X}) \neq 0,\, \delta_{ij} > o_{ij}\}$, $\mathbb{T}(\mathbf{O}, \mathbf{X}) = \{(i,j) : i \neq j,\, d_{ij}(\mathbf{X}) = 0,\, \delta_{ij} > o_{ij}\}$ and $\mathbb{Q}(\mathbf{O}, \mathbf{X}) = \{(i,j) : i = j,\, \delta_{ij} > o_{ij}\}$. The iterations in (3) and (4) start with a randomly chosen initial configuration $\mathbf{X}^{(0)}$ and a zero initial outlier matrix $\mathbf{O}^{(0)}$.

$\mathbf{O}^{(t+1)}$ estimation, via (3), is an ℓ_1 regularization (LASSO) problem. Despite the attenuation of outliers in (2), there is still susceptibility to gross errors, since (4) is a LS solution of the minimization of $\left\| \mathbf{L}\mathbf{X}^{(t+1)} - \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)} \right\|_F^2$. Thus, we propose to a) substitute the aforesaid Frobenius norm with an M-estimator by passing the residual $\mathbf{L}\mathbf{X} - \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$ from a non-negative and differentiable function $\phi(\cdot)$ with respect to (w.r.t.) $\mathbf{X}$ and b) impose a regularization term through the nuclear norm of $\mathbf{X}$, i.e.,

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \{\phi(\mathbf{L}\mathbf{X} - \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}) + \lambda_2 \|\mathbf{X}\|_*\}. \quad (6)$$

M-estimators replace the LS loss function by another that increases less than the squared error [13] and are inaugurated in order to attain supplementary resilience to inaccurate estimation of $\mathbf{O}^{(t+1)}$. The nuclear norm regularization term is introduced in (6) in order to avert the over-smoothness of the Frobenius norm employed in [1].

3. AN HQ FRAMEWORK FOR MDS WITH GROSS ERRORS

In this section, (6) is solved via HQ minimization [14]. Let us rewrite (6) as $\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}}\{\phi(\mathbf{X}) + h(\mathbf{X})\}$ where $h(\mathbf{X}) = \lambda_2 \|\mathbf{X}\|_*$. The potential function can be expressed as $\phi(\mathbf{X}) = \min_{\mathbf{P}}\{Q(\mathbf{X}, \mathbf{P}) + \psi(\mathbf{P})\}$ $\forall \mathbf{X} \in \mathbb{R}^{N\times d}$ where $\mathbf{P} \in \mathbb{R}^{N\times d}$ is the matrix of auxiliary variables, $Q(\mathbf{X}, \mathbf{P})$ is a quadratic function for any $\mathbf{P}$, and $\psi(\cdot)$ is the conjugate function of $\phi(\cdot)$ [15, ch. 3, p. 90]. In particular, $\psi(\mathbf{P}) = \sum_{n=1}^{N}\sum_{m=1}^{d} \psi(p_{nm})$. Thus, we solve for

$$(\hat{\mathbf{X}}, \hat{\mathbf{P}}) = \underset{\mathbf{X},\mathbf{P}}{\operatorname{argmin}}\{J(\mathbf{X}, \mathbf{P})\} = \underset{\mathbf{X},\mathbf{P}}{\operatorname{argmin}}\Big\{ Q(\mathbf{X}, \mathbf{P}) + \sum_{n=1}^{N}\sum_{m=1}^{d} \psi(p_{nm}) + h(\mathbf{X}) \Big\}. \quad (7)$$

An iterative solution for $(\hat{\mathbf{X}}, \hat{\mathbf{P}})$ is given by:

$$\mathbf{P}^{(t+1)} = \delta(\mathbf{X}^{(t)}) \quad (8)$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\{Q(\mathbf{X}, \mathbf{P}^{(t+1)}) + h(\mathbf{X})\}. \quad (9)$$

The HQ minimization admits two forms, namely the multiplicative form and the additive one. To begin with, let us deal with scalar p and x. For $p \in \mathbb{R}_+$ and $x \in \mathbb{R}$, $Q(x, p)$ is a quadratic function defined as $Q_M(x, p) = px^2$ in the multiplicative form. The resulting potential function is $\phi(x) = \min_{p}\{p\,x^2 + \psi(p)\}$ [16]. In the additive form, for $p \in \mathbb{R}$ and $x \in \mathbb{R}$, $Q_A(x, p) = (x\sqrt{c} - \frac{p}{\sqrt{c}})^2$ [17], resulting to the potential function $\phi(x) = \min_{p}\{(x\sqrt{c} - \frac{p}{\sqrt{c}})^2 + \psi(p)\}$, where c is a positive constant with its optimal value estimated by $c = \sup_{x\in\mathbb{R}} \phi''(x)$ [14]. The minimizer function $\delta(\cdot)$ for $\phi(x)$: $\mathbb{R} \to \mathbb{R}$ in the additive and multiplicative forms is defined as [14]:

$$\delta_A(x) = cx - \phi'(x) \quad (10)$$

$$\delta_M(x) = \begin{cases} \phi''(0^+) & \text{if } x = 0 \\ \frac{\phi'(x)}{x} & \text{if } x \neq 0. \end{cases} \quad (11)$$

The auxiliary variables p_{nm} in (8) are determined componentwise by the minimizer function $\delta(\cdot)$ associated to $\phi(\cdot)$. Potential functions $\phi(x) : \mathbb{R} \to \mathbb{R}$ for various M-estimators and their corresponding minimizer functions $\delta(x) : \mathbb{R} \to \mathbb{R}$ for both forms can be found in [1].

3.1. Multiplicative Form (HQMMDSNN)

Next, the just described HQ framework is extended to matrix and vector arguments of the quadratic function $Q_M(\cdot)$, which is defined as: $Q_M(\mathbf{L}\mathbf{X} - \mathbf{L}_{+}\mathbf{X}^{(t)}, \mathbf{p}) = \sum_{i=1}^{N} p_i \left\| (\mathbf{L}\mathbf{X} - \mathbf{L}_{+}\mathbf{X}^{(t)})^i \right\|_2^2$, where $\mathbf{p} \in \mathbb{R}^{N\times 1}$ denotes the vector of the auxiliary variables controlled by the minimizer function $\delta_M(\cdot)$ defined in (11). Then, the potential loss function is defined as $\phi_M(\mathbf{L}\mathbf{X} - \mathbf{L}_{+}\mathbf{X}^{(t)}) = \min_{\mathbf{p}}\{\sum_{i=1}^{N} p_i \left\| (\mathbf{L}\mathbf{X} - \mathbf{L}_{+}\mathbf{X}^{(t)})^i \right\|_2^2 + \sum_{i=1}^{N} \psi(p_i)\}$. Let $\mathbf{Y} = \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$. The objective function takes the form:

$$J_M(\mathbf{X}, \mathbf{p}) = \sum_{i=1}^{N} p_i \left\| (\mathbf{L}\mathbf{X} - \mathbf{Y})^i \right\|_2^2 + \sum_{i=1}^{N} \psi(p_i) + \lambda_2 \|\mathbf{X}\|_*. \quad (12)$$

Let $(\hat{\mathbf{X}}, \hat{\mathbf{p}}) = \underset{\mathbf{X},\mathbf{p}}{\operatorname{argmin}}\{J_M(\mathbf{X}, \mathbf{p})\}$. When $\mathbf{X}$ is sought, the terms $\psi(\cdot)$ are omitted, since the auxiliary variables in (8) are contingent on the minimizer function $\delta_M(\cdot)$ and are fixed. Finally, a local minimizer $(\hat{\mathbf{X}}, \hat{\mathbf{p}})$ can be determined by the alternating minimization:

$$p_i^{(t+1)} = \delta_M\left(\left\| (\mathbf{L}\mathbf{X}^{(t)} - \mathbf{Y})^i \right\|_2 \right) \quad (13)$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\Big\{ \operatorname{tr}((\mathbf{L}\mathbf{X} - \mathbf{Y})^T \mathbf{P}^{(t+1)} (\mathbf{L}\mathbf{X} - \mathbf{Y})) + \lambda_2 \|\mathbf{X}\|_* \Big\} \quad (14)$$

where $\mathbf{P}^{(t+1)} = \text{diag}(\mathbf{p}^{(t+1)})$ is a diagonal matrix with ii-th element equal to $p_i^{(t+1)}$. Setting the derivative of (14) w.r.t. $\mathbf{X}$ equal to zero, a closed-form solution is obtained:

$$\mathbf{X}^{(t+1)} = (\mathbf{L}^T \mathbf{P}^{(t+1)} \mathbf{L})^{-1} (\mathbf{L}^T \mathbf{P}^{(t+1)} \mathbf{Y} - \frac{\lambda_2}{2} \mathbf{U}^{(t)} (\mathbf{V}^{(t)})^T) \tag{15}$$

where $\mathbf{X}^{(t)} = \mathbf{U}^{(t)} \mathbf{\Sigma}^{(t)} (\mathbf{V}^{(t)})^T$ is the SVD of $\mathbf{X}^{(t)}$. For $\mathbf{X}^{(t)} \in \mathbb{R}^{N \times d}$, $\mathbf{U}^{(t)} \in \mathbb{R}^{N \times d}$, $\mathbf{\Sigma}^{(t)} \in \mathbb{R}^{d \times d}$, and $\mathbf{V}^{(t)} \in \mathbb{R}^{d \times d}$. The auxiliary variable p_i represents the weight that controls the impact of $\left\| (\mathbf{L}\mathbf{X} - \mathbf{L}_+ \mathbf{X}^{(t)})^i \right\|_2$. The objective function $J_M(\mathbf{X}, \mathbf{p})$ is curtailed at each iteration until convergence due to the HQ minimization. The employment of M-estimators tapers off the outliers repercussion, since $p_i^{(t+1)}$ admits a low weight due to $\delta_M(\cdot)$ definition in (13) associated with an M-estimator potential function $\phi_M(\cdot)$.

3.2. Additive Form (HQAMDSNN)

For the additive form, a similar alternating minimization procedure can be obtained:

$$\mathbf{P}^{(t+1)} = \delta_A(\mathbf{L}\mathbf{X}^{(t)} - \mathbf{Y}) \tag{16}$$

$$\mathbf{X}^{(t+1)} = (\mathbf{L}^T\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{H}^{(t+1)} - \frac{\lambda_2}{2c} \mathbf{U}^{(t)} (\mathbf{V}^{(t)})^T) \tag{17}$$

where $\mathbf{H}^{(t+1)} = \mathbf{Y} + \frac{1}{c} \mathbf{P}^{(t+1)}$. In both forms, the initial outlier matrix $\mathbf{O}^{(0)}$ is set to zero, the initial embedding $\mathbf{X}^{(0)}$ is chosen randomly, while each entry of $\mathbf{O}^{(t+1)}$ at each iteration is estimated via (3).

4. NUMERICAL TESTS

The performance of the proposed algorithms was benchmarked against three well known MDS techniques implemented in the same environment and tested on the same matrices. These techniques were: a) the SMACOF algorithm [3], b) the REE algorithm in its subgradient version [10], and c) the RMDS [4].

The embedding quality of each algorithm has been appraised w.r.t. the following figures of merit: a) the normalized outlier-free stress defined as $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}}) = \sqrt{\frac{\sum_{(i,j)\in\mathbb{U}}(\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{(i,j)\in\mathbb{U}} \delta_{ij}^2}}$, where $\mathbb{U}$ declares the set of outlier-free dissimilarities ($[\mathbf{O}]_{ij} = 0$), as in [4]; b) the estimated number of outliers $\hat{S}$, as in [4]; c) the raw stress $\sigma_r(\hat{\mathbf{X}})$, defined in (1), between the final embedding and the outlier-free configuration; and d) the standardized Procrustean goodness-of-fit criterion ϱ, applied only to fixed configurations, defined as the squared errors sum standardized by a measure of the scale $\mathbf{X}$[1].

In order to judge algorithms performance, 100 Monte Carlo simulations of RMDS took place with a different random initial configuration $\mathbf{X}^{(0)}$ in each run. The run where RMDS raw stress $\sigma_r(\hat{\mathbf{X}})$ acquired its minimum value was adopted. The RMDS, HQAMDSNN, and HQMMDSNN algorithms terminated if $\left\| \mathbf{X}^{(t+1)} - \mathbf{X}^{(t)} \right\|_F / \left\| \mathbf{X}^{(t+1)} \right\|_F$ was less than 10^{-6} or when the number of iterations attained 5000.

4.1. Cross Data

The first set is a fixed configuration of $N = 65$ points in the two-dimensional space, arranged in a cross. In any branch, the sixteen points are equidistant by one unit. The center of the cross is at (16, 16). Each element of the initial dissimilarity matrix $\mathbf{\Delta}$ was contaminated with a background noise ϵ_{ij}, derived from a zero mean truncated Gaussian distribution with variance $\sigma^2 = 0.1$ and range $[-1, 1]$, in order to avert negative values in $\mathbf{\Delta}$. The outliers indices were chosen randomly, with outliers values being derived from a uniform distribution in $[0, 3 \max \delta_{ij}]$. The percentage of the gross error corruption ϖ was set at 10%.

Let a_h declare the Huber M-estimator parameter and $\hat{\sigma}_\epsilon$ denote the median absolute deviation (MAD)[2] of nominal errors. Imposing the equivalence with Huber M-estimator ($\lambda_1 = 2a_h$) [18] and implementing $a_h = 1.345 \times 1.483 \times \hat{\sigma}_\epsilon$ which yields 95% asymptotic efficiency for the Gaussian distribution [19], λ_1 was set to $3.99\,\hat{\sigma}_\epsilon = 0.8492$ for RMDS and the proposed algorithms. The figures of merit related to the embedding quality delivered by the SMACOF, REE and RMDS are collected in Table 1. Due to the lack of space, only the raw stress $\sigma_r(\hat{\mathbf{X}})$ of the proposed algorithms is plotted in Figure 1 for $\lambda_2 \in [1, 100]$. In particular for the HQMMDSNN, a is set to 10^3, 0.25, 0.25 for the Welsch, Fair and log-cosh M-estimators, respectively. The plots of $\sigma_r(\hat{\mathbf{X}})$ for the additive form and the Fair and Huber M-estimators with $a = 8$ are also overlaid. It can be seen that $\sigma_r(\hat{\mathbf{X}})$ for HQMMDSNN and the Welsch M-estimator and HQAMDSNN with the Fair and Huber M-estimators coincide. Moreover, the proposed algorithms outperform the state-of-the-art MDS techniques for a wide range of λ_2 values w.r.t. $\sigma_r(\hat{\mathbf{X}})$. The plots of ϱ for the proposed algorithms and the aforementioned M-estimators are roughly the same with those of $\sigma_r(\hat{\mathbf{X}})$ and are always smaller than RMDS for $\lambda_2 \in [1, 100]$. The estimated number of outliers $\hat{S}$ is relatively constant, admitting values in the range $[504, 512]$. For most values of $\lambda_2 \in [1, 100]$, the proposed algorithms admit smaller $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ values than the RMDS. However, they demonstrated an unstable performance, implying that $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ alone without $\sigma_r(\hat{\mathbf{X}})$ is not a reliable figure of merit for assessing algorithms' performance.

The range $[1, 100]$, shown in Fig. 1, constitutes a small interval of values for λ_2 used to evaluate the performance of the proposed algorithms. For instance, HQMMDSNN employing the Welsch M-estimator with $\lambda_1 = 0.8492$ and $a = 10^3$ exhibits a better performance than RMDS w.r.t. $\sigma_r(\hat{\mathbf{X}})$ for $\lambda_2 \in [1, 2827]$. Furthermore, the choice $\lambda_2 = 1416$ yields the minimum raw stress configuration ($\sigma_r(\hat{\mathbf{X}}) = 19.95$) from all integer λ_2 values. It is worth noting that SMACOF yields $\sigma_r(\hat{\mathbf{X}}) = 12.727$ on non-contaminated data.

Table 1: Figures of merit for the embedding quality obtained by SMACOF, REE, and RMDS applied to cross data.

Outlier percentage $\varpi = 10\%$	SMACOF	REE	RMDS
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.6866	0.7250	0.0158
Estimated outliers $\hat{S}$	-	-	511
Procrustean goodness-of-fit ϱ	0.7965	0.0003	0.00017
Raw Stress $\sigma_r(\hat{\mathbf{X}})$	141916	38.136	26.619

4.2. Scholastic Aptitude Test Data

The second data set entails real data from average Scholastic Aptitude Test (SAT) scores for the $N = 51$ states in the US, including

[1] $sum(sum((\mathbf{X} - repmat(mean(\mathbf{X}, 1), size(\mathbf{X}, 1), 1)).^2, 1))$.

[2] Median of the absolute deviations of nominal errors from their median.

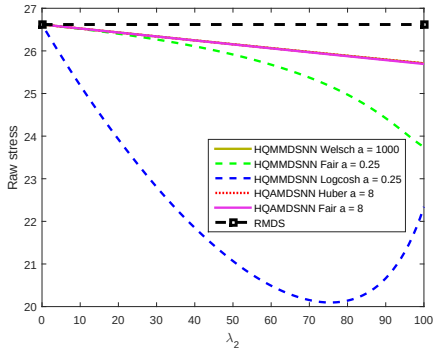

Fig. 1: Raw stress $\sigma_r(\hat{\mathbf{X}})$ of the proposed algorithms for cross data.

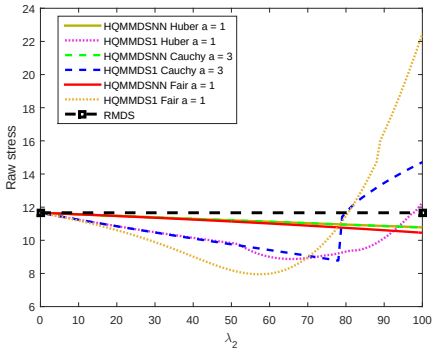

Fig. 2: HQMMDSNN and HQMMDS1 [1] raw stress $\sigma_r(\hat{\mathbf{X}})$ for SAT data.

six attributes, such as population, average verbal and math scores, percentage of eligible students taking the exam, percentage of adult population without a high school education, and annual teacher pay in thousands of dollars [20]. The minimum value of each attribute was subtracted from the initial values of the corresponding attribute and the resulting value was divided by the difference between the maximum and the minimum of each attribute, in order to normalize the initial values in $[0, 1]$. Then, the dissimilarity matrix was computed. The data set was artificially contaminated by $128/(51 \cdot 50/2) = 10.04\%$ outliers drawn from a uniform distribution in $[\max \delta_{ij}, 4 \max \delta_{ij}]$. The outliers indices were chosen randomly. λ_1 was set to 0.75 in order to identify $\hat{S} = 128$ outliers in RMDS with the same value being used in HQMMDSNN. The figures of merit for SMACOF, REE, and RMDS algorithms are gathered in Table 2. HQMMDSNN was proven to be faster than HQAMD-

Table 2: Figures of merit for the embedding quality obtained by SMACOF, REE, and RMDS applied to SAT data.

Outlier percentage $\varpi = 10.04\%$	SMACOF	REE	RMDS
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.6862	0.7608	0.1511
Estimated outliers $\hat{S}$	-	-	128
Raw stress $\sigma_r(\hat{\mathbf{X}})$	251.317	11.7846	11.6615

SNN. Accordingly, only the HQMMDSNN will be considered. For comparison purposes, we include also the performance of HQMMDS1 [1]. Due to space limitations, we shall confine ourselves to $\sigma_r(\hat{\mathbf{X}})$ only. Figure 2 depicts $\sigma_r(\hat{\mathbf{X}})$ for $\lambda_2 \in [1, 100]$. The parameter a was set to 3, 1, and 1 for the Cauchy, Huber, and Fair M-estimators, respectively. The superior performance of the nuclear norm is attributed to the mitigation of the over-smoothness imposed by the Frobenius norm. Hence, for the same parameter a, the range of λ_2 values where the proposed algorithms accomplish a smaller $\sigma_r(\hat{\mathbf{X}})$ than RMDS is always greater than that obtained by HQMMDS1 and HQAMDS algorithms proposed in [1].

4.3. Discussion

To determine if a given dissimilarity matrix $\mathbf{\Delta}$ is corrupted with gross errors, SMACOF and the proposed algorithms are applied for $\lambda_2 = 0$. If the raw stress $\sigma_r(\hat{\mathbf{X}})$ estimated by SMACOF is smaller than that of the proposed algorithms, then the dissimilarity matrix is error-free. The proposed algorithms outperform the state-of-the-art competing techniques, delivering a more accurate approximation of the real configuration for a wide range of λ_2 values.

M-estimator selection: Welsch, Cauchy, Fair, and Huber M-estimators are found to be more stable than others.

Parameter selection: The order of parameter selection is: following order: λ_1, a, λ_2 for the multiplicative form and λ_1, c, a, λ_2 for the additive one. If the MAD of the nominal errors σ_ϵ is accessible, then $\lambda_1 = 3,99\sigma_\epsilon$. Alternatively, the plot of $\hat{S}$ versus λ_1 for the RMDS algorithm is employed by choosing the λ_1 value where this plot exhibits an elbow. The typical choice of the parameter c is $c = \phi''(0)$. The kernel size a of Welsch, Cauchy and Fair M-estimators in both forms may be selected by $\hat{a}^2 = \frac{\left\|\mathbf{L}\mathbf{X}^{(0)} - \mathbf{L}_+\mathbf{X}^{(0)}\right\|_F^2}{2Nd}$ [21] or by applying Silverman's rule [22]. Let $\hat{a}$ be the kernel size determined by the aforementioned rules. A rule of thumb is $a = \xi\hat{a}$ for $\xi \in [2, 4]$.

Unavailability of the error-free dissimilarity matrix: One of the proposed algorithms is implemented for a reasonable range of λ_2 values, with λ_1 being selected according to the elbow rule. Next, the configuration with minimum $\hat{S}$ value is chosen. The latter is proven to be near to that with the minimum $\sigma_r(\hat{\mathbf{X}})$. If $\hat{S}$ is stable for this λ_2 values range, then the embedding with the minimum $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ is selected.

Computational time and complexity: The multiplicative form, incorporating alternating updates of $\mathbf{O}$, $\mathbf{P}$ and $\mathbf{X}$ and SVD costs $O(N^3)$ per iteration in the worst case. The same applies for the additive form. Even though the additive and the multiplicative forms solve the same HQ optimization problem, the multiplicative form exhibits two principal advantages compared to the additive one: a) It requires fewer iterations to converge; b) The tuning of parameter a is found to be more simple.

5. CONCLUSIONS

Two algorithms have been proposed for solving MDS in the presence of gross errors in the dissimilarity matrix. They have been found to outperform the state-of-the-art MDS techniques. The estimation of the optimum λ_2 parameter could be addressed in future research.

6. REFERENCES

[1] F. Mandanas and C. Kotropoulos, "Robust multidimensional scaling using a maximum correntropy criterion," *IEEE Trans. Signal Processing*, vol. 65, no. 4, pp. 919–932, 2017.

[2] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[3] J. de Leeuw, "Applications of convex analysis to multidimensional scaling," in *Recent Developments in Statistics*, J.R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, Eds., pp. 133–146. North Holland, Amsterdam, The Netherlands, 1977.

[4] P. A. Forero and G. B. Giannakis, "Sparsity-exploiting robust multidimensional scaling.," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4118–4134, 2012.

[5] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[6] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2008.

[7] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, 2011.

[8] N. Feiping, H. Heng, and D. Chris, "Low-rank matrix recovery via efficient schatten p-norm minimization," in *Proc. 26th AAAI Conf. Artificial Intelligence*, 2012, pp. 655–661.

[9] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC, 2015.

[10] L. Cayton and S. Dasgupta, "Robust Euclidean embedding," in *Proc. 23rd Int. Conf. Machine Learning*, 2006, pp. 169–176.

[11] W. J. Heiser, "Multidimensional scaling with least absolute residuals," in *Proc. 1st Conf. Int. Federation of Classification Societies (IFCS)*, Aachen, Germany, 1987, pp. 455–462.

[12] W. J. Heiser, *Notes on the LARAMP Algorithm*, Internal Report, Department of Data Theory. University of Leiden, 1987.

[13] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Processing*, vol. 55, no. 11, pp. 5286–5298, 2007.

[14] M. Nikolova and M. K. Ng, "Analysis of half-quadratic minimization methods for signal and image recovery," *SIAM J. Scientific Computing*, vol. 27, no. 3, pp. 937–966, 2005.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[16] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[17] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[18] J.-J. Fuchs, "An inverse problem approach to robust regression," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Washington DC, USA, 1999, vol. 4, pp. 1809–1812.

[19] I. Pitas and A. N. Venetsanopoulos, *Nonlinear Digital Filters: Principles and Applications*, vol. 84, The Springer International Series in Engineering and Computer Science, 1990.

[20] "Stats, statistical datasets," `http://people.sc.fsu.edu/~jburkardt/datasets/stats/stats.html`, Accessed June 11, 2014.

[21] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, "Correntropy induced ℓ_2 graph for robust subspace clustering," in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 1801–1808.

[22] R. He, B.-G. Hu, W.-S. Zheng, and X. Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Trans. Image Processing*, vol. 20, no. 6, pp. 1485–1494, 2011.