M-estimators for Robust Multidimensional Scaling employing $\ell_{2,1}$ norm regularization

Fotios Mandanas*, Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, 54124, Greece

Abstract

Multidimensional Scaling (MDS) has been exploited to visualise the hidden structures among a set of entities in a reduced dimensional metric space. Here, we are interested in cases whenever the initial dissimilarity matrix is contaminated by outliers. It is well-known that the state-of-the-art algorithms for solving the MDS problem generate erroneous embeddings due to the distortion introduced by such outliers. To remedy this vulnerability, a unified framework for the solution of MDS problem is proposed, which resorts to half-quadratic optimization and employs potential functions of M-estimators in combination with $\ell_{2,1}$ norm regularization. Two novel algorithms are derived. Their performance is assessed for various M-estimators against state-of-the-art MDS algorithms on four benchmark data sets. The numerical tests demonstrate that the proposed algorithms perform better than the competing alternatives.

Keywords: Multidimensional scaling, robustness, M-estimators, half-quadratic optimization, $\ell_{2,1}$ norm regularization

1. Introduction

Multidimensional Scaling (MDS) seeks for a visual representation of proximities among a set of entities so that the distances between the entities in the

*Corresponding author: Tel.:+30-231-099-8225; fax: +30-231-099-8473;
Email addresses: fmandan@gmail.com (Fotios Mandanas), costas@aiia.csd.auth.gr (Constantine Kotropoulos)

low-dimensional reconstructed map preserve the initial pairwise dissimilarities as closely as possible. MDS input is a square, symmetric dissimilarity matrix that captures the proximities among a set of entities. Its output is a configuration (i.e., a geometric model), where each entity is represented by a single point. The spectrum of MDS applications includes, psychology [1], construction of market structure maps [2], dimensionality reduction [3], graph drawing [4], phone caller network visualization [5], localization of mobile phones [6], unraveling relational patterns among genes [7], localizing nodes in a wireless sensor network [8], and open-domain sentiment analysis [9].

Widely used MDS algorithms, such as the classical MDS [1] and the scaling by majorizing a complicated function (SMACOF) [10], do not exhibit robustness when the initial dissimilarities are corrupted with outliers. This assumption and the work in [11] have motivated us to propose a variant of the framework presented in [12]. The major premise is that by employing M-estimators to mitigate the repercussion of outliers, contaminating the dissimilarity matrix, and imposing an $\ell_{2,1}$ norm regularization for smoothness, the aforementioned vulnerability of state-of-the-art MDS algorithms is alleviated. Accordingly, the contributions of the paper are: 1) The proposal of a general framework for the solution of the MDS problem when the initial dissimilarity matrix is corrupted with outliers, which is based on half-quadratic (HQ) optimization. 2) The detailed demonstration of the merits of the proposed algorithms against the state-of-the-art MDS algorithms and the study of the impact of the $\ell_{2,1}$ norm regularization against the Frobenius norm used in [12].

The motivation behind using the $\ell_{2,1}$ norm regularization stems from the related literature in statistics and machine learning, when this norm is used as both loss function and regularization term. In particular, the $\ell_{2,1}$ norm has been initially applied in Group Least Absolute Shrinkage and Selection Operator (LASSO) [13], multi-task feature learning [14], [15] and logistic group LASSO [16]. In [17], the subspace learning problem is reformulated using the $\ell_{2,1}$ norm of the projection matrix. A feature selection method imposing joint $\ell_{2,1}$ norm minimization on both the loss function and the regularization term is

proposed in [18]. The $\ell_{2,1}$ norm as a loss function offers robustness to outliers as opposed to the ℓ_2 norm. By employing this norm as a regularization term, sparsity-promoting feature selection is achieved. The $\ell_{2,1}$ norm as a regularization term was proposed for joint embedding learning and sparse regression [19], discriminative feature selection for unsupervised learning [20], and robust feature selection [21]. In [22], a feature selection technique that employs the $\ell_{2,1}$ norm regularization into the Fisher criterion was proposed. In all cases, the $\ell_{2,1}$ norm constraint assures row-sparsity of the feature selection matrix, which leads to informative features.

This paper is structured as follows: Notation and norm definitions are summarized in Section 2. MDS and existing robust variants of MDS are presented in Section 3. The proposed MDS algorithms, employing M-estimators and $\ell_{2,1}$ norm regularization, are detailed in Section 4. Their performance is compared to that of state-of-the-art MDS algorithms by numerical tests in Section 5. Section 6 concludes the paper and proposes topics of future research.

2. Preliminaries

Scalars appear as lowercase letters (e.g., λ_1) while vectors and matrices are denoted by lowercase boldface letters (e.g., $\mathbf{x}$) and uppercase ones (e.g., $\mathbf{X}$), respectively. The (i, j) element of $\mathbf{X}$ is declared as $[\mathbf{X}]_{ij}$ or x_{ij}. The i-th row of $\mathbf{X}$ is represented by the row vector $\mathbf{x}^i$ while the j-th column by the column vector $\mathbf{x}_j$. $(\cdot)^T$ denotes transposition, $\mathbf{I}$ stands for the identity matrix with compatible dimensions, $\mathrm{tr}(\mathbf{X})$ refers to the trace of matrix $\mathbf{X}$, and $\mathbf{X}^{-1}$ is the inverse of the square matrix $\mathbf{X}$. The operator $\mathrm{diag}(\cdot)$ applied to vector $\mathbf{x}$ yields a square diagonal matrix whose main diagonal elements are the elements of $\mathbf{x}$. When the same operator is applied to matrix, i.e., $\mathrm{diag}(\mathbf{X})$ yields a column vector with elements, the ones appearing on the main diagonal of $\mathbf{X}$. The expression $\sum_{i<j}^{N}(\cdot)$ is a short-hand notation for the double summation $\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\cdot)$, while $|\cdot|$ denotes the absolute value operator.

In this paper, we deal with vector and matrix norms. The ℓ_p norm of $\mathbf{x} \in$

$\mathbb{R}^{d\times 1}$ is defined as $\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d}|x_i|^p\right)^{\frac{1}{p}}$. Special cases are the ℓ_1 and ℓ_2 norms of $\mathbf{x}$, equal to $\|\mathbf{x}\|_1 = \sum_{i=1}^{d}|x_i|$ and $\|\mathbf{x}\|_2 = \sqrt{\sum_i^d x_i^2}$, respectively.

Let $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|, ..., |\mathbf{x}_N]^T \in \mathbb{R}^{N\times d}$ be the data matrix, where the i-th object is mapped to the i-th column of $\mathbf{X}^T$ or the i-th row of $\mathbf{X}$, i.e., $\mathbf{x}^i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{id}) \in \mathbb{R}^{1\times d}$. The Frobenius norm of $\mathbf{X} \in \mathbb{R}^{N\times d}$ is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{d} x_{ij}^2} = \sqrt{\sum_{i=1}^{N}\|\mathbf{x}^i\|_2^2}$. The $\ell_{p,q}$ norm of $\mathbf{X}$[1] is defined as:

$$\|\mathbf{X}\|_{p,q} = \left(\sum_{i=1}^{N}\left(\sum_{j=1}^{d}|x_{ij}|^p\right)^{\frac{q}{p}}\right)^{\frac{1}{q}} = \left(\sum_{i=1}^{N}\left\|\mathbf{x}^i\right\|_p^q\right)^{\frac{1}{q}}. \tag{1}$$

For $p = 2$ and $q = 1$, the $\ell_{2,1}$ norm of $\mathbf{X}$ results, i.e., $\|\mathbf{X}\|_{2,1} = \sum_{i=1}^{N}\left\|\mathbf{x}^i\right\|_2$. For any rotation matrix $\mathbf{R}$, it can be proven that $\|\mathbf{X}\,\mathbf{R}\|_{2,1} = \|\mathbf{X}\|_{2,1}$. Moreover, the $\ell_{2,1}$ norm of $\mathbf{X}$ satisfies the three norm conditions.

3. Robust variants of MDS

Let N be the number of objects, d be the resulting embedding dimension, and $\mathbf{\Delta} = [\delta_{ij}]$ denote the initial pairwise dissimilarity matrix, where δ_{ij}, $i, j = 1, 2, \ldots, N$ stands for the dissimilarity between the objects i and j. The embedding in the d-dimensional space is represented by $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|, ..., |\mathbf{x}_N]^T \in \mathbb{R}^{N\times d}$. Let $\mathbf{D}(\mathbf{X}) = [d_{ij}(\mathbf{X})] \in \mathbb{R}^{N\times N}$ denote the distance matrix. Its ij-th element is the ℓ_2 norm between $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e., $d_{ij}(\mathbf{X}) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. The Hadamard product of $\mathbf{D}(\mathbf{X})$ with itself is equal to

$$[\mathbf{D}(\mathbf{X})]^2 = \operatorname{diag}(\operatorname{diag}(\mathbf{X}\mathbf{X}^T))\,\mathbf{E} + \mathbf{E}\operatorname{diag}(\operatorname{diag}(\mathbf{X}\mathbf{X}^T)) - 2\mathbf{X}\mathbf{X}^T \tag{2}$$

where $\mathbf{E}$ is a $N \times N$ matrix whose all elements equal to one. The MDS aims at determining $\mathbf{X}$ so that the raw stress

$$\sigma_r(\mathbf{X}) = \sum_{i<j}^{N}(\delta_{ij} - d_{ij}(\mathbf{X}))^2 \tag{3}$$

[1] Another closely related norm is the so-called ℓ_2/ℓ_1 norm, which is defined as the sum of the ℓ_2 norms of the column vectors of $\mathbf{X}$ and used in joint sparse and low-rank representations [23]. The latter is wrongly defined as $\ell_{2,1}$ in Wikipedia.

is minimized. The raw stress (3) is a least-squares (LS) loss function. Accordingly, it is fragile to outliers. This fragility has led to the proposal of robust Euclidean embedding (REE) [24], where the function $\left\|\boldsymbol{\Delta}^2 - \mathbf{D}^2\right\|_1$ was employed in order to minimize the influence of outliers. Closely related ideas can be found in [25, 26].

Another robust variant is the RMDS [11], where each dissimilarity is modeled as $\delta_{ij} = d_{ij}(\mathbf{X}) + o_{ij} + \epsilon_{ij}$, where o_{ij} denotes an outlier and ϵ_{ij} is a zero-mean independent random variable modeling the nominal error. Since only a small amount of outliers admit a non-zero value, the inclusion of the ℓ_1 norm of the $N \times N$ outlier matrix $\mathbf{O}$ in the MDS loss function is justified, yielding [11]:

$$(\hat{\mathbf{O}}, \hat{\mathbf{X}}) = \underset{\mathbf{O},\mathbf{X}}{\operatorname{argmin}} \left\{ \sum_{i<j}^{N} (\delta_{ij} - d_{ij}(\mathbf{X}) - o_{ij})^2 + \lambda_1 \sum_{i<j}^{N} |o_{ij}| \right\}. \tag{4}$$

The solution of (4) is given by the iterative procedure [11]:

$$o_{ij}^{(t+1)} = S_{\lambda_1}(\delta_{ij} - d_{ij}(\mathbf{X}^{(t)})) \tag{5}$$

$$\mathbf{X}^{(t+1)} = \mathbf{L}^{\dagger}\, \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\, \mathbf{X}^{(t)} \tag{6}$$

where $S_{\lambda_1}(x) = \operatorname{sign}(x)(|x| - \frac{\lambda_1}{2})_+$ is the soft-thresholding operator with $(\cdot)_+ = \max\{\cdot, 0\}$. $\mathbf{L}$ is a symmetric matrix with diagonal elements $[\mathbf{L}]_{ii} = N - 1$ and off-diagonal elements $[\mathbf{L}]_{ij} = -1$. Since $\mathbf{L}$ is not full rank, the Moore-Penrose pseudoinverse is used, defined as $\mathbf{L}^{\dagger} = N^{-1}\mathbf{J}$, 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. In (6), $\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})$ is the Laplacian matrix, i.e.:

$$[\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})]_{ij} = \begin{cases} -(\delta_{ij} - o_{ij})\, d_{ij}^{-1}(\mathbf{X}) & (i,j) \in \mathbb{S}(\mathbf{O}, \mathbf{X}) \\ 0 & (i,j) \in \mathbb{T}(\mathbf{O}, \mathbf{X}) \\ -\sum_{k=1, k\neq i}^{N} [\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})]_{ik} & (i,j) \in \mathbb{Q}(\mathbf{O}, \mathbf{X}) \end{cases} \tag{7}$$

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 iterative procedure starts with a randomly chosen initial configuration $\mathbf{X}^{(0)}$ and a zero initial outlier matrix $\mathbf{O}^{(0)}$.

Given $\mathbf{X}^{(t)}$, the estimation of $\mathbf{O}^{(t+1)}$ via (5) constitutes an ℓ_1 regularization LASSO problem. Given $\mathbf{O}^{(t)}$, the estimation of $\mathbf{X}^{(t+1)}$ via (6) is the LS solution of the optimization problem $\left\|\mathbf{L}\mathbf{X}^{(t+1)} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}\right\|_F^2$.

Even though (4) alleviates the impact of the outliers, it is still vulnerable to them, because (6) is a LS solution strongly influenced by outliers. Here, it is proposed to replace the squared Frobenius norm that yields (6) with an M-estimator by passing the residual $\mathbf{L}\mathbf{X} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$ through a non-negative and differentiable function $\phi(\cdot)$ with respect to (w.r.t.) $\mathbf{X}$, known as *potential function*. Moreover, a smoothness regularization term is simultaneously imposed through the $\ell_{2,1}$ norm of $\mathbf{X}$, i.e.,

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ \phi(\mathbf{L}\mathbf{X} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}) + \lambda_2 \left\|\mathbf{X}\right\|_{2,1} \right\}. \tag{8}$$

M-estimators substitute the LS loss function, being fragile to outliers, with a potential function $\phi(\cdot)$, which expands less than the LS one [27]. By this way, the vulnerability to gross errors is mitigated. The properties of potential functions can be found in [28]. Here, the M-estimator seeks to attenuate the impact of outlying residual errors offering additional immunity to inaccurate estimation of $\mathbf{O}^{(t+1)}$. The $\ell_{2,1}$ norm regularization term combines the benefits of ℓ_1 and ℓ_2 regularization, avoiding the over-smoothness of the Frobenius norm used in [12]. By doing so, we overcome the crucial limitations of RMDS, namely the LS viewpoint that yields (6) and the lack of any smoothness term w.r.t. $\mathbf{X}$.

4. MDS employing M-estimators and $\ell_{2,1}$ regularization

In this section, (8) is solved via half-quadratic (HQ) minimization. To do so, a new objective function is introduced that is more tractable. It depends on both the initial variables $\mathbf{X}$ and new auxiliary variables $\mathbf{P}$. Let $\mathcal{J}(\mathbf{X})$ be the initial objective function and $J(\mathbf{X}, \mathbf{P})$ be the new objective function. These objective functions satisfy $\mathcal{J}(\mathbf{X}) = \min_{\mathbf{P}}\{J(\mathbf{X}, \mathbf{P})\}$, $\forall \mathbf{X}$. That is, the global minimum of $J(\mathbf{X}, \mathbf{P})$ with w.r.t. $\mathbf{X}$ is the same with that of the $\mathcal{J}(\mathbf{X})$. When $\mathbf{P}$ is fixed, J is quadratic w.r.t. $\mathbf{X}$, a property which is responsible for the term *Half Quadratic*

(HQ). In principle, HQ minimization resorts to alternating estimates of $\mathbf{P}$ and $\mathbf{X}$. The optimization problem (8) can be rewritten as

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}}\{\phi(\mathbf{X}) + h(\mathbf{X})\} \tag{9}$$

where $h(\mathbf{X}) = \lambda_2 \left\|\mathbf{X}\right\|_{2,1}$. The potential (or loss) function $\phi(\mathbf{X})$: $\mathbb{R} \to \mathbb{R}$ could be either convex or non-convex and can be chosen to be the potential function of an M-estimator. It also satisfies

$$\phi(\mathbf{X}) = \min_{\mathbf{P}}\{Q(\mathbf{X}, \mathbf{P}) + \psi(\mathbf{P})\} \quad \forall \mathbf{X} \in \mathbb{R}^{N \times d}. \tag{10}$$

In (10), $\mathbf{P} \in \mathbb{R}^{N \times d}$ is a matrix of auxiliary variables, $Q(\mathbf{X}, \mathbf{P})$ is a quadratic function for any $\mathbf{P}$, while $\psi(\cdot)$ is the conjugate function of $\phi(\cdot)$ [29, ch. 3, p. 90]. For $\mathbf{P} \in \mathbb{R}^{N \times d}$, $\psi(\mathbf{P}) = \sum_{n=1}^{N} \sum_{m=1}^{d} \psi(p_{nm})$. By combining (9) and (10), we arrive at

$$\begin{aligned}(\hat{\mathbf{X}}, \hat{\mathbf{P}}) = \underset{\mathbf{X},\mathbf{P}}{\operatorname{argmin}}\{J(\mathbf{X}, \mathbf{P})\} = \underset{\mathbf{X},\mathbf{P}}{\operatorname{argmin}}\Big\{ & Q(\mathbf{X}, \mathbf{P}) \\ & + \sum_{n=1}^{N} \sum_{m=1}^{d} \psi(p_{nm}) + h(\mathbf{X})\Big\}.\end{aligned} \tag{11}$$

The solution $(\hat{\mathbf{X}}, \hat{\mathbf{P}})$ of the optimization problem (11) is obtained in an alternating fashion as follows:

$$\mathbf{P}^{(t+1)} = \delta(\mathbf{X}^{(t)}) \tag{12}$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\{Q(\mathbf{X}, \mathbf{P}^{(t+1)}) + h(\mathbf{X})\}. \tag{13}$$

The auxiliary variables p_{nm} in (12) are determined componentwise by the HQ minimizer function $\delta(\cdot)$ derived by $\psi(\cdot)$ and thus related to $\phi(\cdot)$. Therefore, it suffices to consider the case of scalar functions in (10). For a scalar variable, the minimizer function satisfies the constraint $Q(x, \delta(x)) + \psi(\delta(x)) \leq Q(x, p) + \psi(p)$, $\forall p \in \mathbb{R}$ [30]. $Q(x, p)$ is a quadratic function admitting two forms, namely the multiplicative form

$$Q_M(x, p) = px^2 \qquad p \in \mathbb{R}_+, \quad x \in \mathbb{R} \tag{14}$$

resulting to the potential function $\phi(x) = \min_p\{p\,x^2 + \psi(p)\}$ [31] and the additive form [32]:

$$Q_A(x,p) = (x\sqrt{c} - \frac{p}{\sqrt{c}})^2 \qquad p \in \mathbb{R}, \quad x \in \mathbb{R} \tag{15}$$

which results to the potential function $\phi(x) = \min_p\{(x\sqrt{c} - \frac{p}{\sqrt{c}})^2 + \psi(p)\}$, where c is a positive constant. $c = \sup_{x\in\mathbb{R}} \phi''(x)$ is the optimal value of c [30]. In both forms of the HQ minimization, $\phi(x)$ should fulfil certain conditions [30].

The minimizer function $\delta(\cdot)$, called also *weighting function*, admits distinct additive and multiplicative formulations. For $\phi(x)$: $\mathbb{R} \to \mathbb{R}$, these formulations are [30]:

$$\delta_A(x) = cx - \phi'(x) \tag{16}$$

$$\delta_M(x) = \begin{cases} \phi''(0^+) & \text{if } x = 0 \\ \dfrac{\phi'(x)}{x} & \text{if } x \neq 0. \end{cases} \tag{17}$$

Various potential functions $\phi(x) : \mathbb{R} \to \mathbb{R}$ and their corresponding weighting functions $\delta(x) : \mathbb{R} \to \mathbb{R}$ can be derived from (16) and (17) for the additive and multiplicative form of the HQ, respectively. They are listed in Table 1. In the following, closed forms for (12) and (13) are derived for the additive and multiplicative forms of HQ.

Table 1: Potential functions $\phi(x)$ and weighting functions $\delta(x)$ of M-estimators for either the additive or multiplicative form of HQ.

M-estimator	***Potential Function***	***Multiplicative Form***	***Additive Form***
ℓ_2	$\phi(x) = x^2/2$	$\delta(x) = 1$	$\delta(x) = (c-1)x$
ℓ_1	$\phi(x) = \lvert x\rvert$	$\delta(x) = \frac{1}{\lvert x\rvert}$	$\delta(x) = cx - \text{sign}(x)$
ℓ_p	$\phi(x) = \frac{\lvert x\rvert^p}{p}$ $p \in (1,2]$	$\delta(x) = \lvert x\rvert^{p-2}$	Not Applicable
ℓ_1-ℓ_2	$\phi(x) = 2(\sqrt{1+\frac{x^2}{2}} - 1)$	$\delta(x) = \frac{1}{\sqrt{1+\frac{x^2}{2}}}$	$\delta(x) = cx - \frac{x}{\sqrt{1+\frac{x^2}{2}}}$
Log-cosh	$\phi(x) = \log(\cosh ax)$	$\delta(x) = a\frac{\tanh ax}{x}$	$\delta(x) = cx - a\tanh ax$
Huber	$\phi(x) = \begin{cases} x^2/2 & \lvert x\rvert \le a \\ a\lvert x\rvert - \frac{a^2}{2} & \lvert x\rvert > a \end{cases}$	$\delta(x) = \begin{cases} 1 & \lvert x\rvert \le a \\ \frac{a}{\lvert x\rvert} & \lvert x\rvert > a \end{cases}$	$\delta(x) = \begin{cases} (c-1)x & \lvert x\rvert \le a \\ cx - a\text{sign}(x) & \lvert x\rvert > a \end{cases}$
Fair	$\phi(x) = a^2(\frac{\lvert x\rvert}{a} - \log(1+\frac{\lvert x\rvert}{a}))$	$\delta(x) = \frac{1}{1+\frac{\lvert x\rvert}{a}}$	$\delta(x) = cx - \frac{x}{1+\frac{\lvert x\rvert}{a}}$
Welsch	$\phi(x) = \frac{a^2}{2}(1 - \exp(-\frac{x^2}{a^2}))$	$\delta(x) = \exp(-\frac{x^2}{a^2})$	$\delta(x) = cx - x\exp(-\frac{x^2}{a^2})$
Cauchy	$\phi(x) = \frac{a^2}{2}\log(1 + (\frac{x}{a})^2)$	$\delta(x) = \frac{1}{1+(\frac{x}{a})^2}$	$\delta(x) = cx - \frac{x}{1+(\frac{x}{a})^2}$
Geman-McClure	$\phi(x) = \frac{x^2}{2(1+x^2)}$	$\delta(x) = \frac{1}{(1+x^2)^2}$	$\delta(x) = cx - \frac{x}{(1+x^2)^2}$
Tukey	$\phi(x) = \begin{cases} \frac{a^2}{6}(1 - [1 - (\frac{x}{a})^2]^3) & \lvert x\rvert \le a \\ \frac{a^2}{6} & \lvert x\rvert > a \end{cases}$	$\delta(x) = \begin{cases} [1 - (\frac{x}{a})^2]^2 & \lvert x\rvert \le a \\ 0 & \lvert x\rvert > a \end{cases}$	$\delta(x) = \begin{cases} cx - x[1 - (\frac{x}{a})^2]^2 & \lvert x\rvert \le a \\ 0 & \lvert x\rvert > a \end{cases}$

4.1. Additive Form (HQAMDSL21)

By adapting (15) to the multivariate case, the quadratic function $Q_A(\cdot)$ of the additive form of the HQ is defined as

$$Q_A(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}, \mathbf{P}) = \left\| \sqrt{c}\,(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) - \frac{1}{\sqrt{c}}\,\mathbf{P} \right\|_F^2 \tag{18}$$

where the matrix of auxiliary variables $\mathbf{P} \in \mathbb{R}^{N\times d}$ is determined by the minimizer function $\delta_A(\cdot)$ defined in (16). Hence, the potential loss function $\phi_A(\cdot)$, applying (10), takes the form:

$$\begin{aligned}\phi_A(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) = \min_{\mathbf{P}}\Bigg\{ & Q_A(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}, \mathbf{P}) \\ & + \sum_{n=1}^{N}\sum_{m=1}^{d} \psi(p_{nm}) \Bigg\}\end{aligned} \tag{19}$$

where $\psi(\cdot)$ is the conjugate function of $\phi_A(\cdot)$. Accordingly, $J_A(\mathbf{X}, \mathbf{P})$ in (11) is given by:

$$\begin{aligned}J_A(\mathbf{X}, \mathbf{P}) = & \left\| \sqrt{c}\,(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) - \frac{1}{\sqrt{c}}\,\mathbf{P} \right\|_F^2 \\ & + \sum_{n=1}^{N}\sum_{m=1}^{d} \psi(p_{nm}) + \lambda_2 \left\|\mathbf{X}\right\|_{2,1}\end{aligned} \tag{20}$$

where λ_2 is a positive parameter regulating the $\ell_{2,1}$ norm of $\mathbf{X}$. Let $(\hat{\mathbf{X}}, \hat{\mathbf{P}}) = \underset{\mathbf{X},\mathbf{P}}{\operatorname{argmin}}\{J_A(\mathbf{X}, \mathbf{P})\}$. When $\mathbf{X}$ is sought, the terms including $\psi(\cdot)$ can be omitted. Recall that the auxiliary variables depend only on the minimizer function $\delta_A(\cdot)$, as indicated in (12), and are fixed. Let $\mathbf{Y} = \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$. Then, the unknown variables $(\mathbf{X}, \mathbf{P})$ are estimated by the alternating minimization procedure:

$$\mathbf{P}^{(t+1)} = \delta_A(\mathbf{L}\mathbf{X}^{(t)} - \mathbf{Y}) \tag{21}$$

$$\begin{aligned}\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\Bigg\{ & \left\| \sqrt{c}\,(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) - \frac{1}{\sqrt{c}}\,\mathbf{P}^{(t+1)} \right\|_F^2 \\ & + \lambda_2 \left\|\mathbf{X}\right\|_{2,1} \Bigg\}.\end{aligned} \tag{22}$$

The optimization problem (22) can be reformulated as:

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \Big\{ c \operatorname{tr}((\mathbf{L}\mathbf{X} - \mathbf{H}^{(t+1)})^T(\mathbf{L}\mathbf{X} - \mathbf{H}^{(t+1)})) + 2\lambda_2 \operatorname{tr}(\mathbf{X}^T\mathbf{R}^{(t+1)}\mathbf{X}) \Big\}. \tag{23}$$

where

$$\mathbf{H}^{(t+1)} = \mathbf{Y} + \frac{1}{c}\,\mathbf{P}^{(t+1)}. \tag{24}$$

and

$$\mathbf{R}_{ii}^{(t+1)} = r_i^{(t+1)} = \frac{1}{2\left\|(\mathbf{x}^i)^{(t)}\right\|_2} \tag{25}$$

with $\mathbf{R}^{(t+1)} = \operatorname{diag}(\mathbf{r}^{(t+1)})$ being a diagonal matrix with ii-th element equal to $r_i^{(t+1)}$. By applying the first order optimality condition to (23) w.r.t. $\mathbf{X}$, a closed form solution is obtained for $\mathbf{X}^{(t+1)}$[2]:

$$\mathbf{X}^{(t+1)} = c\,(c\mathbf{L}^T\mathbf{L} + \lambda_2\mathbf{R}^{(t+1)})^{-1}\mathbf{L}^T\mathbf{H}^{(t+1)}. \tag{26}$$

In this form of the HQ, the auxiliary variables $\mathbf{P}$ can be viewed as errors incurred by noise. The complete procedure for the solution of (8) by the additive form of the HQ minimization is outlined in Algorithm 1. The initial configuration $\mathbf{X}^{(0)}$ can be chosen randomly, while the initial outlier matrix $\mathbf{O}^{(0)}$ is set to zero.

4.2. Multiplicative Form (HQMMDSL21)

For the multiplicative form of the HQ, the quadratic function $Q_M(\cdot)$ is defined as the weighted sum of squared ℓ_2 norms of the rows of the residual $\mathbf{L}\mathbf{X} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$:

$$Q_M(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}, \mathbf{p}) = \sum_{i=1}^{N} p_i \left\|(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)})^i\right\|_2^2 \tag{27}$$

where $\mathbf{p} \in \mathbb{R}^{N\times 1}$ is the vector of the auxiliary variables, which is determined by the minimizer function $\delta_M(\cdot)$ defined in (17). Thus, the potential loss function

[2]Recall that $\mathbf{L}$ is symmetric, so $\mathbf{L}^T = \mathbf{L}$.

Algorithm 1 Additive form of the HQ Minimization for MDS with $\ell_{2,1}$ regularization (HQAMDSL21)

Input: Initial outlier matrix $\mathbf{O}^{(0)}$ and initial configuration $\mathbf{X}^{(0)}$

Output: Outlier matrix $\mathbf{O}^{(t+1)}$ and coordinate matrix $\mathbf{X}^{(t+1)}$

1: **for** $t = 0, 1, 2, ...$ **do**
2: Find each entry of $\mathbf{O}^{(t+1)}$ via (5)
3: Update $\mathbf{P}^{(t+1)}$ via (21) with $\mathbf{L}_+$ as in (7)
4: Update $\mathbf{H}^{(t+1)}$ via (24) with $\mathbf{L}_+$ as in (7)
5: Update $r_i^{(t+1)}$ via (25)
6: Update $\mathbf{X}^{(t+1)}$ via (26)
7: **end for**

$\phi_M(\cdot)$ is equal to

$$\phi_M(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) = \min_{\mathbf{p}}\Bigg\{\sum_{i=1}^{N} p_i \left\|(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)})^i\right\|_2^2 + \sum_{i=1}^{N}\psi(p_i)\Bigg\}. \tag{28}$$

Using (28), the augmented objective function in (11) takes the form

$$J_M(\mathbf{X}, \mathbf{p}) = \sum_{i=1}^{N} p_i \left\|(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)})^i\right\|_2^2 + \sum_{i=1}^{N}\psi(p_i) + \lambda_2 \left\|\mathbf{X}\right\|_{2,1}. \tag{29}$$

Let $(\hat{\mathbf{X}}, \hat{\mathbf{p}}) = \underset{\mathbf{X},\mathbf{p}}{\operatorname{argmin}}\{J_M(\mathbf{X}, \mathbf{p})\}$. Following similar lines to the derivation of HQAMDSL21, a local minimizer $(\hat{\mathbf{X}}, \hat{\mathbf{p}})$ can be estimated using the alternating minimization:

$$p_i^{(t+1)} = \delta_M\left(\left\|(\mathbf{L}\mathbf{X}^{(t)} - \mathbf{Y})^i\right\|_2\right) \tag{30}$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\Bigg\{\operatorname{tr}((\mathbf{L}\mathbf{X} - \mathbf{Y})^T\mathbf{P}^{(t+1)}(\mathbf{L}\mathbf{X} - \mathbf{Y})) + 2\lambda_2\operatorname{tr}(\mathbf{X}^T\mathbf{R}^{(t+1)}\mathbf{X})\Bigg\} \tag{31}$$

where $\mathbf{P}^{(t+1)} = \text{diag}(\mathbf{p}^{(t+1)})$ and $\mathbf{R}^{(t+1)} = \text{diag}(\mathbf{r}^{(t+1)})$ are diagonal matrices with ii-th elements equal to $p_i^{(t+1)}$ and $r_i^{(t+1)}$ given by (25), respectively. Setting the derivative of (31) w.r.t. $\mathbf{X}$ equal to zero, a closed-form solution is obtained:

$$\mathbf{X}^{(t+1)} = (\mathbf{L}^T \mathbf{P}^{(t+1)} \mathbf{L} + \lambda_2 \mathbf{R}^{(t+1)})^{-1} \mathbf{L}^T \mathbf{P}^{(t+1)} \mathbf{Y}. \tag{32}$$

The HQ minimization guarantees that the objective function is reduced at each iteration until convergence. At each iteration, the auxiliary variable p_i denotes the weight that modulates the influence of $\left\|(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)})^i\right\|_2$. The adoption of M-estimators into the objective function $J_M(\mathbf{X}, \mathbf{p})$ attenuates the outlier impact. This is because $p_i^{(t+1)}$ takes always a low weight, as attested by the $\delta_M(\cdot)$ in (30), which is related to an M-estimator potential function $\phi_M(\cdot)$. The entire procedure for the solution of (8) by the multiplicative form of HQ is outlined in Algorithm 2. $\mathbf{X}^{(0)}$ and $\mathbf{O}^{(0)}$ can be initialized as in Section 4.1.

It should be noted that if $\left\|(\mathbf{x}^i)^{(t)}\right\|_2 = 0$, then $\mathbf{R}_{ii}^{(t+1)} = r_i^{(t+1)}$ will tend to infinity. In this case, there is no guarantee that the proposed algorithms will converge. Although in theory $\left\|\mathbf{x}^i\right\|_2$ can be zero, in practice it should be set to a very small, but non-zero, value. Under these circumstances, $\mathbf{R}_{ii}^{(t+1)}$ can be regularized as $\mathbf{R}_{ii}^{(t+1)} = \frac{1}{2\left\|(\mathbf{x}^i)^{(t)}\right\|_2 + \zeta}$, where ζ is a very small constant. It is true that when $\zeta \to 0$, then $\frac{1}{2\left\|(\mathbf{x}^i)^{(t)}\right\|_2 + \zeta}$ approximates $\frac{1}{2\left\|(\mathbf{x}^i)^{(t)}\right\|_2}$.

5. Numerical Tests

The proposed algorithms were implemented in Matlab and tested on several dissimilarity matrices. Their performance was benchmarked against three state-of-the-art MDS algorithms implemented in the same environment and tested on the same dissimilarity matrices. These algorithms were: a) the SMACOF algorithm [10], b) the subgradient version of REE algorithm [24], and c) the RMDS [11]. For all state-of-the-art algorithms, any authors' recommendations were strictly followed.

The embedding quality of each algorithm has been judged w.r.t. four figures of merit: a) The normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}}) = \sqrt{\frac{\sum_{(i,j)\in\mathbb{U}}(\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{(i,j)\in\mathbb{U}}\delta_{ij}^2}}$

Algorithm 2 Multiplicative form of the HQ minimization for MDS with $\ell_{2,1}$ regularization (HQMMDSL21)

Input: Initial outlier matrix $\mathbf{O}^{(0)}$ and initial configuration $\mathbf{X}^{(0)}$

Output: Outlier matrix $\mathbf{O}^{(t+1)}$ and coordinate matrix $\mathbf{X}^{(t+1)}$

1: **for** $t = 0, 1, 2, \ldots$ **do**
2: Find each entry of $\mathbf{O}^{(t+1)}$ via (5)
3: Update $p_i^{(t+1)}$ via (30) with $\mathbf{L}_+$ as in (7)
4: Update $r_i^{(t+1)}$ via (25)
5: Update $\mathbf{X}^{(t+1)}$ via (32)
6: **end for**

was calculated, where $\mathbb{U}$ denotes the set of outlier-free dissimilarities (i.e., when $[\mathbf{O}]_{ij} = 0$) as in [11]. To calculate this figure of merit, the set of outliers estimated by the applicable algorithm was used. Any ground truth related to the outliers was not employed, so that the figure of merit did not depend on the knowledge of the initial outlier-free dissimilarity matrix. b) The estimated number of outliers $\hat{S}$ was recorded as in [11]. c) The distortion (raw stress) $\sigma_r(\hat{\mathbf{X}})$ between the resulting embedding and the initial outlier-free configuration, defined in (3), was monitored. d) The standardized Procrustean goodness-of-fit criterion ϱ, defined as the sum of the squared errors standardized by a measure of the scale $\mathbf{X}$[3], was calculated, too. The last criterion can only be applied to fixed configurations.

100 Monte Carlo simulations of the RMDS algorithm took place, using a different random initial configuration $\mathbf{X}^{(0)}$ in each run. From all runs, the instance where RMDS embedding was closer to the outlier-free configuration (i.e., when raw stress $\sigma_r(\hat{\mathbf{X}})$ admitted its minimum value) was selected. RMDS, HQAMDSL21, and HQMMDSL21 algorithms terminated if $\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F /$

[3]In Matlab, the measure of the scale $\mathbf{X}$ is given by $sum(sum((\mathbf{X} - repmat(mean(\mathbf{X}, 1), size(\mathbf{X}, 1), 1)).^2, 1))$.

$\left\|\mathbf{X}^{(t+1)}\right\|_F$ was less than 10^{-6} or after 5000 iterations.

The following experimental results demonstrate that if the M-estimators are employed in HQAMDSL21 and HQMMDSL21, the resulting embedding outperforms the one derived by RMDS w.r.t. $\sigma_r(\hat{\mathbf{X}})$ for a wide range of λ_2 values. The same applies for ϱ citerion in fixed configurations. Due to lack of space, only the raw stress $\sigma_r(\hat{\mathbf{X}})$ of the proposed algorithms is plotted for $\lambda_2 \in [1, 100]$.

5.1. Square Data Set

The first data set is a rectangular of $N = 100$ points in the two-dimensional space. The bottom-left point is at $(1, 1)$ and the upper-right point is at $(10, 10)$. All points are equidistant by one unit from their vertical and horizontal neighbors. Each element of the initial dissimilarity matrix was corrupted with a background error ϵ_{ij} extracted from a zero mean truncated Gaussian distribution with variance 0.1 and threshold $-d_{ij}(\mathbf{X})$. This threshold was selected in order to avert negative values in $\mathbf{\Delta}$. The outliers were derived from a uniform distribution in [0,40] with their indices being chosen arbitrarily. The contamination percentage $\varpi\%$ of the outliers was set at $594/(100 \cdot 99/2) = 12\%$.

Let a_h be the parameter of the Huber M-estimator and $\hat{\sigma}_\epsilon$ be the median absolute deviation (MAD)[4] of nominal errors. Implementing the equivalence with Huber M-estimator ($\lambda_1 = 2a_h$) and taking into account that $a_h = 1.345 \times 1.483 \times \hat{\sigma}_\epsilon$ yields 95% asymptotic efficiency for the normal distribution [33], λ_1 was set to $3.99\,\hat{\sigma}_\epsilon$ for the RMDS and the proposed algorithms.

Table 2 gathers the figures of merit related to the embedding quality delivered by SMACOF, REE, and RMDS. The reported figures of REE were recorded after 4000 iterations. The range of the figures of merit for HQMMDSL21, employing the Fair M-estimator with $a = 0.7$ is also included in Table 2 when $\lambda_2 \in [1, 100]$. The raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQAMDSL21 and HQMMDSL21 algorithms is plotted in Figure 1 for $\lambda_2 \in [1, 100]$. In both algorithms, the kernel size

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

Table 2: Figures of merit for the embedding quality obtained by SMACOF, REE, RMDS, and HQMMDSL21 applied to square data set.

Outlier percentage $\varpi = 12\%$	SMACOF	REE	RMDS	HQMMDSL21
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.6830	0.7206	0.0375	[0.0373, 0.03755]
Estimated outliers $\hat{S}$	-	-	1354	[1323, 1359]
Procrustean goodness-of-fit ϱ	0.3925	0.0006	0.0004	[0.00038, 0.0004]
Raw stress $\sigma_r(\hat{\mathbf{X}})$	52728.4	58.0572	51.3491	[34.6436, 51.2819]

a of the Welsch, Cauchy, Fair and log-cosh M-estimators was set to 12, 4, 0.7, and 0.7, respectively. For HQAMDSL21, c was equal to 1. It can be seen that the plots of raw stress $\sigma_r(\hat{\mathbf{X}})$ for HQAMDSL21 and HQMMDSL21, employing the Welsch and Cauchy M-estimator, are rather identical. Furthermore, the figures of merit of the proposed algorithms, employing the Welsch, Cauchy, Fair, and log-cosh M-estimators are summarized in Table 3.

The plots of ϱ for the proposed algorithms and the aforementioned M-estimators are roughly the same with that of $\sigma_r(\hat{\mathbf{X}})$. ϱ for the proposed algorithms is always smaller than that of RMDS for $\lambda_2 \in [1, 100]$. The estimated number of outliers $\hat{S}$ by both HQAMDSL21 and HQMMDSL21 admitted values in [1323, 1359]. $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ for the proposed algorithms takes a smaller value than RMDS for most choices of $\lambda_2 \in [1, 100]$. Nevertheless, the normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ exhibits an unstable performance, indicating that this figure alone without $\hat{S}$ or $\sigma_r(\hat{\mathbf{X}})$ is not reliable for judging the embedding quality.

Finally, in order to demonstrate the influence of the kernel size a of the M-estimators, the plots of the raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQMMDSL21, employing the Fair M-estimator, for different values of a are overlaid in Figure 2, when $\lambda_2 \in [1, 100]$. It can be seen that a large value of the parameter a impedes the derivation of the optimal embedding (i.e., finding the configuration with the lowest raw stress $\sigma_r(\hat{\mathbf{X}})$), which emerges for a large values of λ_2.

Table 3: Range of the figures of merit for the embedding quality obtained by HQMMDSL21 and HQAMDSL21 for various M-estimators applied to square data set.

	$\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	$\hat{S}$	ϱ	$\sigma_r(\hat{\mathbf{X}})$
HQMMDSL21 Welsch $a = 12$	[0.03734, 0.03746]	[1347, 1359]	[0.00039, 0.0004]	[45.178, 51.282]
HQAMDSL21 Welsch $a = 12$	[0.03734, 0.03746]	[1347, 1359]	[0.00039, 0.0004]	[45.180, 51.282]
HQMMDSL21 Cauchy $a = 4$	[0.03734, 0.03746]	[1347, 1359]	[0.00039, 0.0004]	[45.097, 51.282]
HQAMDSL21 Cauchy $a = 4$	[0.03734, 0.03746]	[1347, 1359]	[0.00039, 0.0004]	[45.123, 51.285]
HQMMDSL21 Fair $a = 0.7$	[0.03730, 0.03755]	[1323, 1359]	[0.00038, 0.0004]	[34.644, 51.282]
HQAMDSL21 Fair $a = 0.7$	[0.03722, 0.03747]	[1336, 1359]	[0.00038, 0.0004]	[37.352, 51.284]
HQMMDSL21 Log-cosh $a = 0.7$	[0.03728, 0.03748]	[1334, 1359]	[0.00039, 0.0004]	[37.791, 51.213]
HQAMDSL21 Log-cosh $a = 0.7$	[0.03729, 0.03749]	[1336, 1359]	[0.00038, 0.0004]	[38.374, 51.217]

5.2. Scholastic Aptitude Test Data Set

The second data set comprises real data from average Scholastic Aptitude Test (SAT) scores for the $N = 51$ states in the US. These data include 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 [34]. First, the initial values were normalized in [0,1]. Then, the dissimilarity matrix was computed according to (2). Next, the dissimilarity matrix was artificially contaminated by $128/(51 \cdot 50/2) = 10.04\%$ outliers, being drawn from a uniform distribution in $[\max \delta_{ij}, 4 \max \delta_{ij}]$. The outlier indices were chosen randomly. In order to estimate $\hat{S} = 128$ outliers using the RMDS, λ_1 was set to 0.75. The same value was used for HQMMDSL21.

The figures of merit used to judge the embedding quality obtained by SMA-

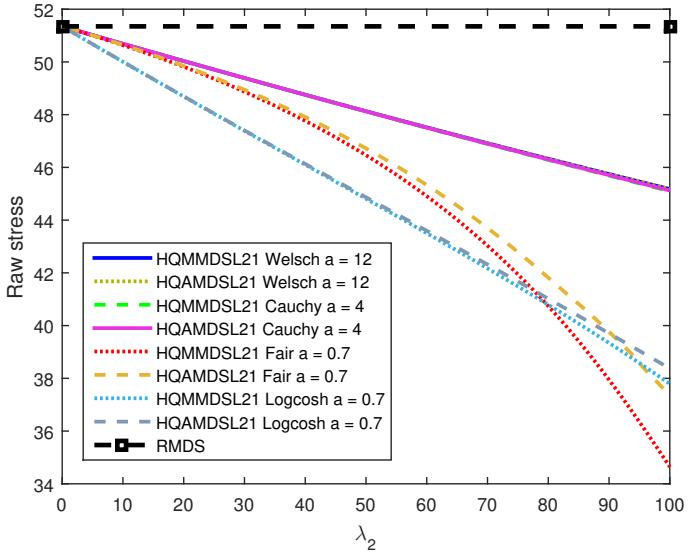


Figure 1: Raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQAMDSL21 and HQMMDSL21 in the square data set.

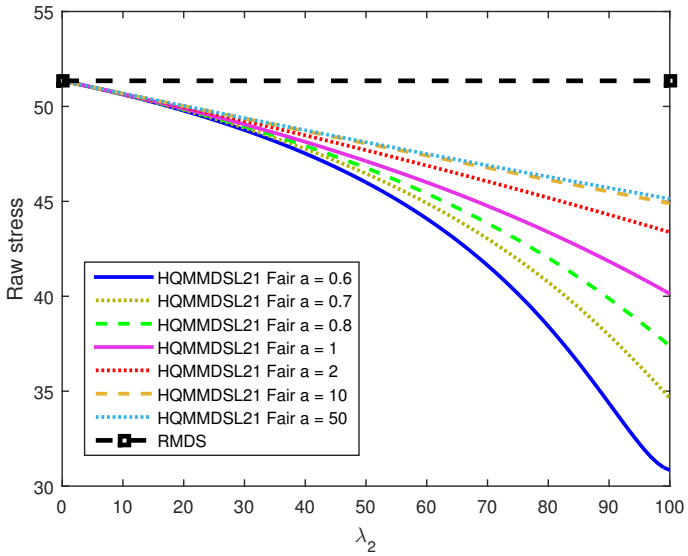


Figure 2: Raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQMMDSL21, employing the Fair M-estimator, for different values of the kernel size a in the square data set.

Table 4: Figures of merit for the embedding quality obtained by SMACOF, REE, RMDS, and HQMMDSL21 applied to SAT data set.

Outlier percentage $\varpi = 10.04\%$	SMACOF	REE	RMDS	HQMMDSL21
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.6862	0.7608	0.1511	[0.138, 0.1508]
Estimated outliers $\hat{S}$	-	-	128	128
Raw stress $\sigma_r(\hat{\mathbf{X}})$	251.3171	11.7846	11.6615	[9.379, 11.623]

COF, REE, and RMDS are summarized in Table 4. The reported figures of REE came after 8000 iterations. For comparison purposes, the range of the figures of merit of the HDMMDSL21, employing the Cauchy M-estimator with a being set to 3, is also incorporated in the same Table, when λ_2 varies in $[1, 100]$. The raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQMMDSL21 for various M-estimators is plotted in Figure 3 for $\lambda_2 \in [1, 100]$.

Detailed figures of merit for HQMMDSL21 employing various M-estimators are listed in Table 5. The estimated number of outliers $\hat{S}$ was found to be

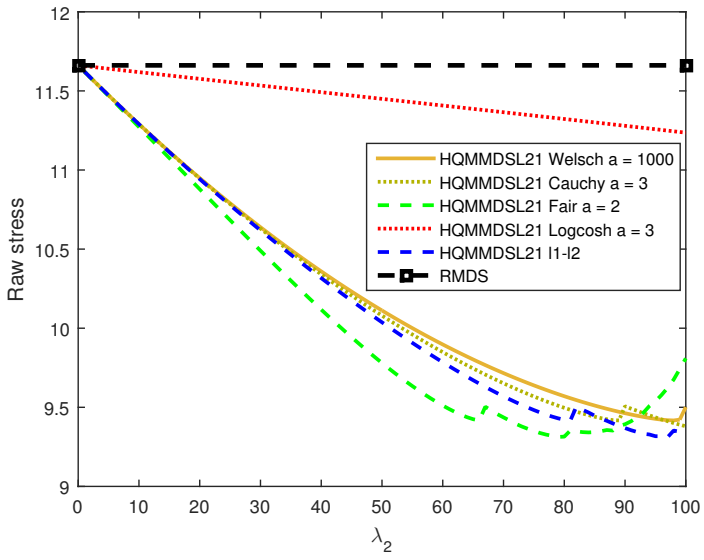


Figure 3: HQMMDSL21 raw stress $\sigma_r(\hat{\mathbf{X}})$ for the SAT data set.

Table 5: Range of figures of merit for the embedding quality obtained by HQMMDSL21 for various M-estimators applied to SAT data set.

	$\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	$\hat{S}$	$\sigma_r(\hat{\mathbf{X}})$
HQMMDSL21 Welsch $a = 1000$	[0.1380, 0.1508]	128	[9.4170, 11.6230]
HQMMDSL21 Cauchy $a = 3$	[0.1380, 0.1508]	128	[9.3797, 11.6230]
HQMMDSL21 Fair $a = 2$	[0.1380, 0.1508]	[128,129]	[9.3128, 11.6228]
HQMMDSL21 Log-cosh $a = 3$	[0.1484, 0.1510]	128	[11.2359, 11.6572]
HQMMDSL21 ℓ_1-ℓ_2	[0.1366, 0.1508]	[128, 129]	[9.0775, 11.6230]

constant, admitting values in the range [128,129]. The normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ of HQMMDSL21 for all M-estimators was always smaller than the same figure of merit of RMDS, when $\lambda_2 \in [1, 100]$. The plot of $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ of HQMMDSL21, employing the aforementioned M-estimator was roughly the same with that of $\sigma_r(\hat{\mathbf{X}})$.

In the reported experiments, the performance of the Welsch M-estimator in HQMMDSL21, with kernel size being selected to a large value ($a = 1000$), approximates that of the ℓ_2 M-estimator in the same algorithm. The value of a above which the aforementioned approximation takes place is different for each M-estimator. In this case, the lowest raw stress value $\sigma_r(\hat{\mathbf{X}})$, derived by HQMMDSL21 algorithm (and HQAMDSL21), is valid for the largest possible λ_2 value. On the contrary, the parameter a of Fair M-estimator with HQMMDSL21 was selected so that the local minimum of the raw stress $\sigma_r(\hat{\mathbf{X}})$ is achieved for a smaller λ_2 value. By this way, the range of λ_2 values for which the proposed algorithms outperform RMDS w.r.t. $\sigma_r(\hat{\mathbf{X}})$ is reduced.

5.3. Cities Data Set

The third data set is composed by the airline distances in hundreds of miles, between $N = 30$ world principal international cities [35]. The initial dissimilarity matrix was articially contaminated by $65/(30 \cdot 29/2) = 14.94\%$ outliers drawn from a uniform distribution in $[0, 3.5 \max \delta_{ij}]$. Outliers indices were selected

randomly. λ_1 was set to 45.63 in order to identify $\hat{S} = 65$ outliers using the RMDS. The same value was used for the proposed algorithms.

Table 6 gathers the figures of merit related to the embedding quality obtained by SMACOF, REE, and RMDS. The reported figures for REE were measured after 20000 iterations. The range of the figures of merit for HQMMDSL21, employing the Fair M-estimator with $a = 3$, is also included in the Table, when $\lambda_2 \in [1, 100]$.

Table 6: Figures of merit for the embedding quality obtained by SMACOF, REE, RMDS, and HQMMDSL21 applied to cities data set.

Outlier percentage $\varpi = 14.94\%$	SMACOF	REE	RMDS	HQMMDSL21
Normalized outlier-free stress $\sigma(\hat{X}, \hat{O})$	0.6039	0.7353	0.1065	[0.1065, 0.1078]
Estimated outliers $\hat{S}$	-	-	65	[64, 65]
Raw stress $\sigma_r(\hat{\mathbf{X}})$	820550	203540	70369	[69326, 70363]

The raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQAMDSL21 and HQMMDSL21 is plotted in Figure 4 for $\lambda_2 \in [1, 100]$. The kernel size a for the Welsch and Cauchy estimators was chosen to be 10^{10} and 30, respectively. For Fair, Huber, and log-cosh M-estimators, a was set to 3. Parameter c was equal to 1 for HQAMDSL21. It is seen that the raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQAMDSL21 and HQMMDSL21, employing the Welsch, Cauchy, and Huber M-estimators, coincide for the selected values of a.

The estimated number of outliers $\hat{S}$, in both forms, was found to be constant, admitting values in [64, 65]. The normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ of the proposed algorithms exhibits larger values than those of RMDS for most choices of λ_2 $[1, 100]$ indicating that $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ alone is not a reliable figure of merit for judging embedding quality.

It is worth noting that the range $\lambda_2 \in [1, 100]$ covers a small portion of the full range of values where the performance of the proposed algorithms has been assessed. For example, the HQMMDSL21, employing the Welsch M-estimator with $\lambda_1 = 45.63$ and $a = 10^{10}$ exhibits a better performance than RMDS w.r.t.

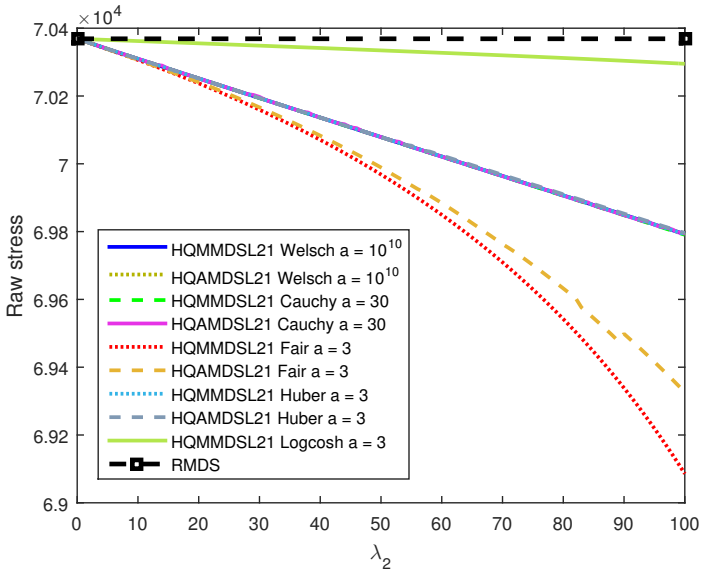
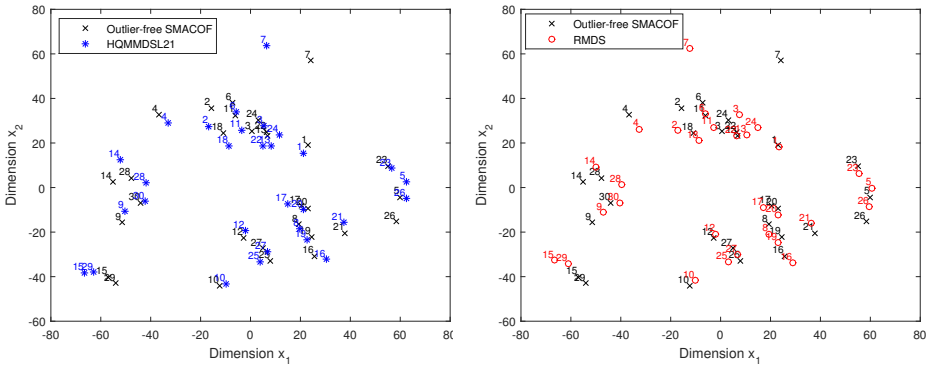


Figure 4: Raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQAMDSL21 and HQMMDSL21 in the cities data set.

$\sigma_r(\hat{\mathbf{X}})$, when $\lambda_2 \in [1, 8631]$. Setting $\lambda_2 = 5305$, we obtained the configuration with the lowest raw stress $\sigma_r(\hat{\mathbf{X}})$ (i.e., 54066) for all integer choices of λ_2.

The embedding of SMACOF applied to outlier-free data was used as baseline to be compared with the embeddings delivered by RMDS and HQMMDSL21, when they are applied to contaminated data. These embeddings are shown in Figures 5a and 5b. The embeddings obtained by RMDS and HQMMDSL21 were matched to that of SMACOF via Procrustes analysis. The embedding of HQMMDSL21 was obtained, when the Welsch M-estimator with kernel size $a = 10^{10}$ was employed, setting $\lambda_1 = 45.63$ and $\lambda_2 = 5305$. A careful visual inspection of Figures 5a and 5b reveals that 12 points derived by HQMMDSL21 are closer to SMACOF benchmark ones than those obtained by RMDS. On the contrary, 9 points obtained by RMDS are closer to SMACOF benchmark, while 9 points of RMDS and HQMMDSL21 algorithms appear to coincide. Nevertheless, the total distance between the corresponding points in HQMMLDSL21 and SMACOF embeddings is smaller than the total distance between the corresponding points in RMDS and SMACOF embeddings. This observation is also



(a) (b)

Figure 5: (a) Embeddings by HQMMDSL21 on contaminated data and SMACOF on outlier-free data; (b) Embeddings by RMDS on contaminated data and SMACOF on outlier free data.

confirmed by $\sigma_r(\hat{\mathbf{X}})$ values: 41498 for SMACOF on outlier-free data, 54066 for HQMMDSL21, and 70369 for RMDS. The last two embeddings were derived from contaminated data. Similarly, the standardized Procrustean goodness-of-fit criterion ϱ w.r.t SMACOF benchmark is 0.0264 for HQMMDSL21 and 0.0463 for RMDS, respectively.

It should be mentioned, however, that the visualization of embeddings leads to less qualitative differences than the estimated figures of merit (e.g., raw stress $\sigma_r(\hat{\mathbf{X}})$ values). This is due to the fact that $\sigma_r(\hat{\mathbf{X}})$ values correspond to the distortion between the resulting embedding and the initial outlier-free configuration. The mapping of SMACOF benchmark, as illustrated in Figures 5a and 5b, corresponds to the best possible approximation of the initial outlier-free configuration, which may diverge significantly from the real one. Furthermore, when a data set consists of many points, less apparent differences between the embeddings can be located.

5.4. Packets Data Set

The fourth data set encompasses packet-delay differences derived from a delay-based scheme. This scheme involves three packets sent from a fixed source.

That is, a small packet is first sent to terminal node i followed by a large packet sent to node j and finally a small packet is sent once more to node i [36]. The network includes $N = 10$ terminal nodes, generating $(10 \cdot 9)/2 = 45$ terminal pairs. Each measurement stems from the difference between the arrival times of the first and second small packet at terminal node i, being relevant to the path bandwidth shared with terminal node j [36]. The scheme was implemented $9,567$ times totally, incorporating swaps between the small and the large packet terminal nodes.

The mean packet-delays τ_{ij}, denoting path similarities, represent the non-contaminated (outlier free) data. These data are transformed into dissimilarities via $\delta_{ij} = 100 \exp(-\frac{\tau_{ij}}{1000})$, as in [11]. For each terminal pair, the same formulation was imposed on minimum and maximum packet delays to acquire their largest δ_{ij}^{max} and smallest δ_{ij}^{min} dissimilarities, respectively [11]. The data was artificially contaminated by 12 outliers, drawn from a uniform distribution in $[\delta_{ij}^{min}, \delta_{ij}^{max}]$. The outliers indices were chosen randomly. λ_1 was set to 29.9 in order to identify $\hat{S} = 12$ outliers with RMDS. The same parameter value was used in HQMMDSL21.

The figures of merit related to the embedding quality of SMACOF, REE, and RMDS algorithms are gathered in Table 7. The reported figures of REE were recorded after 4000 iterations. The range of figures of merit of HQMMDSL21, employing the Welsch M-estimator with kernel size a set to 1000, is also included in Table 7, when λ_2 varies in $[1, 100]$. Taking into account that

Table 7: Figures of merit for the embedding quality obtained by SMACOF, REE, RMDS, and HQMMDSL21 applied to packets data set.

Outlier percentage $\varpi = 26.67\%$	SMACOF	REE	RMDS	HQMMDSL21
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.3222	0.4706	0.1755	[0.1714, 0.1808]
Estimated outliers $\hat{S}$	-	-	12	[11, 13]
Raw stress $\sigma_r(\hat{\mathbf{X}})$	22345.02	5941.1	14111.4	[12159.1, 14087.9]

the multiplicative form was found to be faster than the additive one, only $\sigma_r(\hat{\mathbf{X}})$

for HQMMDSL21 is plotted in Figure 6 for $\lambda_2 \in [1, 100]$. Parameter a was set to 1000 for the Welsch, Cauchy, and Fair M-estimators. The plots of $\sigma_r(\hat{\mathbf{X}})$ for the HQMMDS1 algorithm [12], employing the aforementioned M-estimators with identical kernel size a are overlaid in Figure 6.

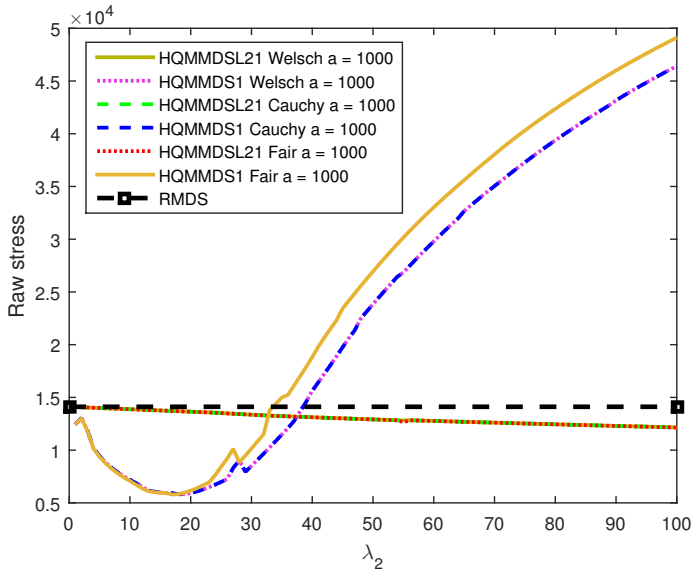


Figure 6: Raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQMMDSL21 and HQMMDS1[12] in the packets data set.

It is seen that HQMMDSL21 outperforms RMDS w.r.t. $\sigma_r(\hat{\mathbf{X}})$ for $\lambda_2 \in [1, 100]$ with the plots of Welsch, Cauchy and Fair M-estimators to coincide. On the contrary, the raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQMMDS1, employing either the Welsch or Cauchy M-estimator, whose plots are superimposed, admits smaller values than those obtained by RMDS for $\lambda_2 \in [1, 38]$. When the Fair M-estimator is employed in HQMMDSL21, the values of λ_2 guaranteeing a better performance than RMDS are limited to $[1, 33]$. The range of the figures of merit of HQMMMDSL21 and HQMMDS1 [12], employing various M-estimators is summarized in Table 8.

The replacement of the Frobenius norm used in [12] with the $\ell_{2,1}$ norm, yields a completely different performance. This can be attributed to the fact that $\ell_{2,1}$ norm regularization, being in between the ℓ_1 norm and the Frobe-

Table 8: Range of the figures of merit for the embedding quality obtained by HQMMDSL21 and HQMMDS1 [12] with various M-estimators applied to packet data set.

	$\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	$\hat{S}$	$\sigma_r(\hat{\mathbf{X}})$
HQMMDSL21 Welsch $a = 1000$	[0.1714, 0.1808]	[11, 13]	[12159.1, 14087.9]
HQMMDSL21 Cauchy $a = 1000$	[0.1714, 0.1808]	[11, 13]	[12152.6, 14087.9]
HQMMDSL21 Fair $a = 1000$	[0.1713, 0.1808]	[11, 13]	[12144.2, 14087.9]
HQMMDS1 Welsch $a = 1000$	[0.1581, 0.5290]	[10, 33]	[5816.2, 46417.9]
HQMMDS1 Cauchy $a = 1000$	[0.1581, 0.5290]	[10, 33]	[5816.2, 46415.9]
HQMMDS1 Fair $a = 1000$	[0.1581, 0.5149]	[10, 37]	[5786.4, 49115.2]

nius one, mitigates the over-smoothness imposed by the Frobenius norm. For the same kernel size a, the range of λ_2 values for which HQMMDSL21 and HQAMDSL21 attain a smaller raw stress $\sigma_r(\hat{\mathbf{X}})$ than RMDS, is always wider than that for HQMMDS1 and HQAMDS algorithms proposed in [12]. The minimum value of raw stress $\sigma_r(\hat{\mathbf{X}})$ achieved by HQMMDSL21 and HQMMDS1, when the Welsch M-estimator is employed, is 5176.7 (for $\lambda_2 = 618$) and 5816.18 (for $\lambda_2 = 18$), respectively. HQMMDSL21 exhibited a rather unstable performance w.r.t. $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$. For the same algorithm, $\hat{S}$ was proven to be relatively constant, admitting values in [11,13].

The embedding derived by SMACOF on outlier-free data and those derived by HQMMDSL21 and RMDS on contaminated data are illustrated in Figure 7. HQMMDSL21 and RMDS embeddings have been matched to that of SMACOF via Procrustes analysis. The HQMMDSL21 embedding was obtained by employing the Welsch M-estimator with $\lambda_1 = 29.9$, $\lambda_2 = 618$, and $a = 1000$. An attentive examination of Figure 7 discloses that six points $(1, 3, 6, 7, 9, 10)$ in HQMMDSL21 embedding are closer to SMACOF benchmark than the corresponding points derived by RMDS. On the contrary, four points $(2, 4, 5, 8)$ obtained by RMDS are closer to SMACOF benchmark. However, comparing the aforementioned embeddings carefully, it is deducted that, as a whole, the embedding derived by HQMMDSL21 is closer to the SMACOF benchmark than

the RMDS one. This is because the total distance between the corresponding points in HQMMLDSL21 and SMACOF embeddings is smaller than that between the corresponding points in RMDS and SMACOF embeddings. This is also validated by $\sigma_r(\hat{\mathbf{X}})$ values: 2222,7 for SMACOF on outlier-free data; 5176.3 for HQMMDSL21 and 14111.4 for RMDS on contaminated data. The standardized Procrustean goodness-of-fit criterion ϱ w.r.t. SMACOF benchmark is 0.4813 for HQMMDSL21 and 0.5412 for RMDS.

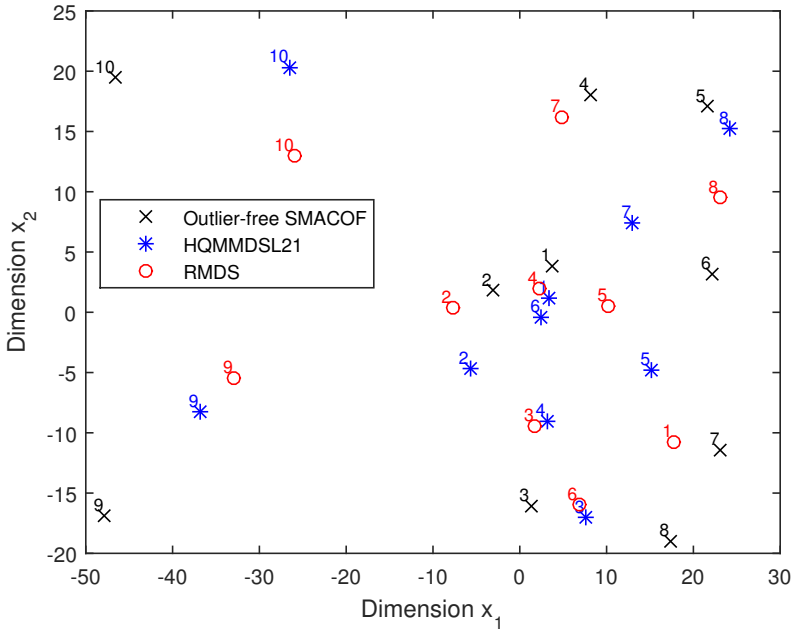


Figure 7: Embeddings obtained by HQMMDSL21 and RMDS on contaminated data juxtaposed to the embedding derived by SMACOF on outlier-free data for packets data set.

Even though the scope of the paper is focused on MDS algorithms, it would be a challenging task to compare the proposed framework with ~~the~~ existing algorithms in visual analytics, ~~e.g., ISOMAP [3], Locally Linear Embedding (LLE) [37], curvilinear component analysis [38], curvilinear distance analysis [39] and~~ t-SNE (t-Stochastic Neighbor Embedding) algorithm [40]. ~~Due to the inherent weaknesses of ISOMAP, LLE, curvilinear component analysis and~~

curvilinear distance analysis, that are discussed briefly in the last Section, t-SNE has been applied to packets dataset. It is worth mentioning that t-SNE supports a dissimilarity matrix as an input. In t-SNE, outliers are circumvented by dening the joint probabilities in the high-dimensional space to be the symmetrized conditional probabilities.

100 Monte Carlo simulations of the t-SNE took place with the same parameters. From all runs, the instance with the lowest value of the ϱ w.r.t. SMACOF benchmark was selected. The embedding derived by SMACOF on outlier-free data and those obtained by HQMMDSL21 and t-SNE on contaminated data, being matched to that of SMACOF via Procrustes analysis, are plotted in Figure 8. It is revealed that six points $(1, 2, 3, 5, 9, 10)$ in HQMMDSL21 embedding are closer to SMACOF benchmark. On the other hand, four points $(4, 6, 7, 8)$ derived by t-SNE are closer to SMACOF benchmark. However, as a whole, the HQMMDSL21 embedding is closer to the SMACOF benchmark than the t-SNE one. Thus, using the same methodology, it is demonstrated that t-SNE cannot accommodate efficiently dissimilarity matrices corrupted with outliers.

5.5. Discussion

The proposed algorithms are shown to perform better than their state-of-the-art competitors. In particular, they yield a better approximation of the true configuration than the RMDS for a wide range of values admitted by λ_2. The REE embedding appears to be considerably better than that of SMACOF, but in most cases the REE embedding is still inferior than that of RMDS. Next, several practical issues are discussed.

M-estimator selection: Extensive experimental results validate that the Welsch, Cauchy, Fair, and Huber M-estimators yield the most stable performance for a wide range of λ_2 values. The efficiency of an M-estimator is determined highly by the proper selection of its kernel size a. The tuning of a for the aforementioned M estimators was found to be easier than the rest M-estimators. The widest range of λ_2 values for which the proposed algorithms attain a smaller raw stress $\sigma_r(\hat{\mathbf{X}})$ than RMDS, is obtained with the ℓ_2 M-estimator.

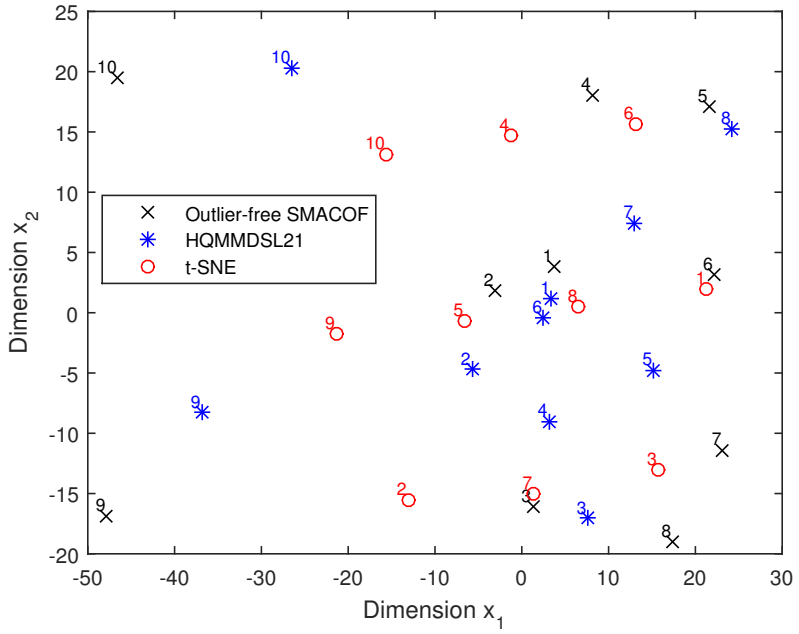


Figure 8: Embeddings obtained by HQMMDSL21 and t-SNE on contaminated data juxtaposed to the embedding derived by SMACOF on outlier-free data for packets data set.

Parameter selection within HQAMDSL21 and HQMMDSL21: The performance of HQMMDSL21 depends on three parameters, namely the regularization weights λ_1 and λ_2 and the kernel size a for each M-estimator. The performance of HQAMDSL21 is also determined by the constant c, which appears in the minimizer function $\delta_A(\cdot)$. For c, the typical choice is $c = \phi''(0)$. The tuning of parameters should be made in the following order: λ_1, a, λ_2 for HQMMDSL21 and λ_1, c, a, λ_2 for HQAMDSL21.

If the MAD of the nominal errors σ_ϵ is available, then $\lambda_1 = 3,99\sigma_\epsilon$ for the Huber M-estimator. Otherwise, the plot of $\hat{S}$ versus λ_1 for the RMDS may be exploited to locate the value of λ_1 for which this curve exhibits an elbow.

The kernel size a of Welsch, Cauchy and Fair M-estimators in both forms can be determined by $a^2 = \frac{\left\|\mathbf{L}\mathbf{X}^{(0)} - \mathbf{L}_+\mathbf{X}^{(0)}\right\|_F^2}{2Nd}$ [41] or by applying Silverman's rule instead [42]. Let $\hat{a}$ be the kernel size estimated by either of the two rules.

A practical recommendation is to set $a = \xi\hat{a}$ for $\xi \in [2, 7]$. Alternatively, one can set the kernel size a of the aforementioned M-estimators much larger than the values predicted in [41] and [42] in order to approximate the performance of the ℓ_2 M-estimator. If a is so tuned, the proposed algorithms yield a more efficient performance than the RMDS for the widest possible range of λ_2 values. In this case, the optimal embedding w.r.t. $\sigma_r(\hat{\mathbf{X}})$ derived by the proposed algorithms takes place for the largest possible λ_2 value. The value of a, above which the equivalence with the ℓ_2 M-estimator is achieved, is different for each M-estimator. It is also highly data-dependent. On the contrary, if the objective is to find the best approximation of the true configuration for a small value of λ_2, then a smaller value of a than that determined in [41] and [42] is recommended. In the latter case, there is a risk of unstable behavior of $\sigma_r(\hat{\mathbf{X}})$ w.r.t. varying λ_2. Moreover, the range of λ_2 values for which the proposed algorithms perform better than RMDS w.r.t. $\sigma_r(\hat{\mathbf{X}})$ is much smaller.

Algorithm comparison: Even though the additive and the multiplicative forms solve the same HQ optimization problem, their performance appears to be different. The tuning of parameter a in the multiplicative form is found to be simpler than in the additive form. Furthermore, HQMMDSL21 requires fewer iterations than HQAMDSL21 to converge, rendering it more appealing than HQAMDSL21 for configurations contaminated with outliers. Thus, it is advised to select a large value of a to achieve stability and then to implement HQMMDSL21. By doing so, fewer iterations are required for convergence.

Unavailability of the outlier-free dissimilarity matrix: Under these circumstances, only $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ and $\hat{S}$ can be used as figures of merit. In this case, HQMMDSL21 or HQAMDSL21 is implemented for a reasonable range of λ_2 values, having selected λ_1 according to the elbow rule. The embedding, which corresponds to the minimum value of $\hat{S}$, is found to be close to that corresponding to the minimum $\sigma_r(\hat{\mathbf{X}})$. If $\hat{S}$ is approximately constant for a range of λ_2 values, then the embedding, which admits the minimum $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$, can be chosen.

Computational time and complexity: In many cases, the proposed algo-

rithms need fewer iterations than the RMDS to converge. The computational complexity of HQMMDSL21, incorporating alternating updates of $\mathbf{O}$, $\mathbf{P}$, $\mathbf{R}$, and $\mathbf{X}$, is $O(N^3)$ per iteration in the worst case. The same applies for the HQAMDSL21 algorithm. That is, their computational complexity per iteration, is the same with the RMDS and the REE. Hence, the proposed algorithms outperform RMDS and REE without increasing the computational complexity. The classical MDS algorithm [1] exhibits $O(N^3)$ total computations, since the eigen-decomposition of a matrix is needed. SMACOF has $O(N^2)$ computational complexity per iteration. It should be accentuated that the number of iterations for the proposed algorithms depends on selection of the M-estimator, its kernel size a, and the regularization parameter λ_2. For their estimation, the proposed algorithms require some additional time as a pre-processing step in order to estimate λ_1, a and λ_2 even though the estimation of a can be avoided by selecting an extremely large value.

Hint: Let $\mathbf{\Delta}$ be a given dissimilarity matrix. To determine if $\mathbf{\Delta}$ has been corrupted by outliers, the SMACOF and one of the proposed algorithms for $\lambda_2 = 0$ can be applied. For an outlier-free dissimilarity matrix, the raw stress $\sigma_r(\mathbf{X})$ of SMACOF has been found to be smaller than that of the proposed algorithms.

6. Conclusions

A new, efficient HQ framework has been proposed for solving the MDS problem when the dissimilarity matrix has been corrupted by outliers. The proposed algorithms have been compared with three state-of-the-art MDS algorithms (i.e., SMACOF, REE, and RMDS) under the same conditions. The experimental findings have manifested that the HQ minimization, in combination with M-estimators and $\ell_{2,1}$ regularization, outperforms the aforementioned competing techniques in either additive or multiplicative form for any configuration being contaminated with outliers. This is of paramount importance in the context of scientific visualization and data mining. It is worth noting that other

variants of the proposed framework could also be explored. For example, the $\ell_{2,1}$ norm can be employed not only for regularization, but in the loss function as well. Moreover, the optimal value of λ_2 can be estimated by solving a proper optimization problem. However, this optimal value of λ_2 will be always greater than that estimated when the Frobenius norm is employed [12]

~~Another topics of~~ future research could ~~be focused on~~ the revision of algorithms used in visual analytics. For example, the ISOMAP algorithm [3] entails three steps, where the last one applies classical MDS, ~~which is extremely prone to outliers.~~ LLE algorithm [37] employs a least squares objective function, which is susceptible to outliers, ~~generating highly corrupted embeddings that diverge considerably from the non-contaminated ones.~~ The cost function used in curvilinear component analysis [38] is quadratic, without exploiting M-estimators or any smoothness term, being used by ~~our~~ proposed algorithms. The same holds for curvilinear distance analysis [39]. ~~To sum up,~~ the proposed algorithms could substitute MDS in ISOMAP, while M-estimators or any smoothness term could be used in the context of LLE, curvilinear component analysis and curvilinear distance analysis. Finally, a variant of the t-SNE algorithm [40] could be developed in order to mitigate more the outliers repercussion.

References

[1] W. S. Torgerson, Multidimensional scaling: I. Theory and method, Psychometrika 17 (4) (1952) 401–419.

[2] J. D. Carroll, P. E. Green, Psychometric methods in marketing research: Part II, Multidimensional scaling, Journal of Marketing Research 34 (2) (1997) 193–204.

[3] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[4] E. R. Gansner, Y. Koren, S. C. North, Graph drawing by stress majorization, in: J. Pach (Ed.), Proc. 12th Int. Conf. Graph Drawing, Vol. LNCS 3383, Springer-Verlag, Berlin, 2005, pp. 239–250.

[5] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, L. Chen, Data visualization with multidimensional scaling, Journal of Computational and Graphical Statistics 17 (2) (2008) 444–472.

[6] K. W. Cheung, H.-C. So, A multidimensional scaling framework for mobile location using time-of-arrival measurements, IEEE Trans. Signal Processing 53 (2-1) (2005) 460–470.

[7] Y. H. Taguchi, Y. Oono, Relational patterns of gene expression via non-metric multidimensional scaling analysis, Bioinformatics 21 (6) (2005) 730–740.

[8] A. Pal, Localization algorithms in wireless sensor networks: Current approaches and future challenges, Network Protocols and Algorithms 2 (1) (2010) 45–74.

[9] E. Cambria, Y. Song, H. Wang, N. Howard, Semantic multi-dimensional scaling for open-domain sentiment analysis, IEEE Intelligent Systems 99 (2) (2014) 44–51.

[10] J. de Leeuw, Applications of convex analysis to multidimensional scaling, in: J. R. Barra, F. Brodeau, G. Romier, B. V. Cutsem (Eds.), Recent Developments in Statistics, North Holland, Amsterdam, The Netherlands, 1977, pp. 133–146.

[11] P. A. Forero, G. B. Giannakis, Sparsity-exploiting robust multidimensional scaling., IEEE Trans. Signal Processing 60 (8) (2012) 4118–4134.

[12] F. Mandanas, C. Kotropoulos, Robust multidimensional scaling using a maximum correntropy criterion, IEEE Trans. Signal Processing 65 (4) (2017) 919–932.

[13] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B 68 (1) (2006) 49–67.

[14] G. Obozinski, B. Taskar, M. Jordan, Multi-task feature selection, Technical Report, Department of Statistics, University of California, Berkeley, 2006.

[15] A. Argyriou, T. Evgeniou, M. Pontil, Multi-task feature learning, in: Proc. 19th Annual Conf. Neural Information Processing Systems, Vol. 19, Vancouver, British Columbia, Canada, 2007, pp. 41–48.

[16] L. Meier, S. Van De Geer, P. Bühlmann, The group LASSO for logistic regression, Journal of the Royal Statistical Society: Series B 70 (2008) 53–71.

[17] Q. Gu, Z. Li, J. Han, Joint feature selection and subspace learning, in: Proc. Int. Joint Conf. Artificial Intelligence, AAAI Press, 2011, pp. 1294–1299.

[18] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization, in: Advances in Neural Information Processing Systems, 2010, pp. 1813–1821.

[19] C. Hou, F. Nie, D. Yi, Y. Wu, Feature selection via joint embedding learning and sparse regression, in: Proc. Int. Joint Conf. Artificial Intelligence, AAAI Press, 2011, pp. 1324–1329.

[20] Y. Yang, H. Shen, Z. Ma, Z. Huang, X. Zhou, ℓ_{21}-norm regularized discriminative feature selection for unsupervised learning, in: Proc. Int. Joint Conf. Artificial Intelligence, 2011, pp. 1589–1594.

[21] R. He, T. Tan, L. Wang, W.-S. Zheng, ℓ_{21} regularized correntropy for robust feature selection, in: Proc. IEEE Computer Vision and Pattern Recognition, 2012, pp. 2504–2511.

[22] J. Zhang, J. Yu, J. Wan, Z. Zeng, ℓ_{21} norm regularized Fisher criterion for optimal feature selection, Neurocomputing 166 (2015) 455–463.

[23] Y. Panagakis, C. Kotropoulos, G. Arce, Music genre classification via joint sparse low-rank representations of audio features, ACM/IEEE Trans. Audio, Speech, and Language Processing 22 (12) (2014) 1905–1915.

[24] L. Cayton, S. Dasgupta, Robust Euclidean embedding, in: Proc. 23rd Int. Conf. Machine Learning, ICML '06, 2006, pp. 169–176.

[25] 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.

[26] W. J. Heiser, Notes on the LARAMP Algorithm, Internal Report RR-87-04, Department of Data Theory, University of Leiden, 1987.

[27] P. J. Huber, Robust estimation of a location parameter, Annals of Mathematical Statistics 55 (1964) 73–101.

[28] W. Liu, P. P. Pokharel, J. C. Principe, Correntropy: Properties and applications in non-Gaussian signal processing, IEEE Trans. Signal Processing 55 (11) (2007) 5286–5298.

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

[30] M. Nikolova, M. K. Ng, Analysis of half-quadratic minimization methods for signal and image recovery, SIAM J. Scientific Computing 27 (3) (2005) 937–966.

[31] D. Geman, G. Reynolds, Constrained restoration and the recovery of discontinuities, IEEE Trans. Pattern Analysis and Machine Intelligence 14 (3) (1992) 367–383.

[32] D. Geman, C. Yang, Nonlinear image recovery with half-quadratic regularization, IEEE Trans. Image Processing 4 (7) (1995) 932–946.

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

[34] Stats, statistical datasets, `http://people.sc.fsu.edu/~jburkardt/datasets/stats/stats.html`, accessed July 10, 2015.

[35] Hartigan, test data for clustering algorithms, `http://people.sc.fsu.edu/~jburkardt/datasets/hartigan/hartigan.html`, accessed June 12, 2014.

[36] M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, Y. Tsang, Maximum likelihood network topology identification from edge-based unicast measurements, SIGMETRICS Perform. Eval. Rev. 30 (1) (2002) 11–20.

[37] S. T. Roweis, L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[38] P. Demartines, J. Herault, Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets, IEEE Trans. Neur. Netw. 8 (1) (1997) 148–154.

[39] J. Lee, A. Lendasse, M. Verleysen, Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis, Neurocomputing 57 (2004) 49–76.

[40] L. van der Maaten, G. E. Hinton, Visualizing high-dimensional data using t-sne, Journal of Machine Learning Research 9 (2008) 2579–2605.

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

[42] R. He, B.-G. Hu, W.-S. Zheng, X. Kong, Robust principal component analysis based on maximum correntropy criterion, IEEE Trans. Image Processing 20 (6) (2011) 1485–1494.