A MAXIMUM CORRENTROPY CRITERION FOR ROBUST MULTIDIMENSIONAL SCALING

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) refers to a class of dimensionality reduction techniques applied to pairwise dissimilarities between objects, so that the interpoint distances in the space of reduced dimensions approximate the initial pairwise dissimilarities as closely as possible. Here, a unified framework is proposed, where the MDS is treated as maximization of a correntropy criterion, which is solved by half-quadratic optimization in a multiplicative formulation. The proposed algorithm is coined as Multiplicative Half-Quadratic MDS (MHQMDS). Its performance is assessed for potential functions associated to various M-estimators, because the correntropy criterion is closely related to the Welsch M-estimator. Three state-of-the-art MDS techniques, namely the Scaling by Majorizing a Complicated Function (SMACOF), the Robust Euclidean Embedding (REE), and the Robust MDS (RMDS), are implemented under the same conditions. The experimental results indicate that the MHQMDS, relying on the M-estimators, performs better than the aforementioned state-of-the-art competing techniques.

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

1. INTRODUCTION

Multidimensional Scaling (MDS) has been widely used for the visualization of hidden structures among the objects in a geometric space. It can be treated as a transformation yielding a geometric model so that the resulting interpoint distances between objects in the new space approximate the initial pairwise dissimilarities as closely as possible. MDS algorithms were firstly inaugurated in psychology [1]. The spectrum of their applications has been expanded to include dimensionality reduction [2], graph drawing [3], phone callers' social network visualization [4], texture mapping on arbitrary surfaces [5], and localization of nodes in a wireless sensor network [6]. Recently, a semantic MDS for open-domain sentiment analysis was developed [7].

However, the traditional techniques for the solution of the MDS problem, like the classical MDS [1] and the scaling by majorizing a complicated function (SMACOF) [8], despite their simplicity, are not robust when the initial dissimilarities are contaminated with outliers. Even a single outlier in the dissimilarity matrix may distort severely the solution of the classical MDS algorithm, because the noise is propagated to each element of the distance matrix through the double-centering process involved [9, 10].

The motivation for this paper stems from the insight that by employing M-estimators in the solution of the MDS problem, robustness to outliers is gained. In particular, when the dissimilarity matrix is contaminated with outliers, the following novel contributions are made: 1) A framework is developed in order to estimate the MDS embedding, that is based on half-quadratic (HQ) minimization in combination with M-estimators; 2) An efficient algorithm for finding the MDS solution is proposed, which is based on the multiplicative form of the HQ; 3) The Welsch M-estimator, which is closely related to the maximum correntropy criterion, is thoroughly studied for solving the MDS problem.

Notation: Scalars are denoted by lowercase letters (e.g., λ_1), vectors appear as lowercase boldface letters (e.g., $\mathbf{x}$), and matrices are indicated by uppercase boldface letters (e.g., $\mathbf{X}$). The ij-th element of $\mathbf{X}$ is represented by $[\mathbf{X}]_{ij}$ or x_{ij}, while $(\,)^T$ denotes transposition. If $\mathbf{X}$ is a square matrix, then $\mathbf{X}^{-1}$ denotes its inverse and $\mathrm{tr}(\mathbf{X})$ is its trace. $\mathbf{I}$ stands for the identity matrix with compatible dimensions, $\mathrm{diag}(\mathbf{x})$ denotes a square diagonal matrix with the elements of vector $\mathbf{x}$ on the main diagonal, while $\mathrm{diag}(\mathbf{X})$ yields a column vector formed by the elements of the main diagonal of $\mathbf{X}$. The i-th row of a matrix $\mathbf{X}$ is declared by the row vector $\mathbf{x}^i$, while the j-th column is indicated with the column vector $\mathbf{x}_j$. If $|\cdot|$ denotes the absolute value operator, then, for $\mathbf{x} \in \mathbb{R}^{n\times 1}$, $\|\mathbf{x}\|_1 = \sum_{i=1}^{n}|x_i|$ and $\|\mathbf{x}\|_2 = \sqrt{\sum_i^n x_i^2}$ are the ℓ_1 and ℓ_2 norms of $\mathbf{x}$, respectively. The Frobenius norm of $\mathbf{X} \in \mathbb{R}^{n\times m}$ is defined as $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2}$.

2. ROBUST MDS APPROACHES

Let N denote the number of objects and d be the embedding dimension. In graphical representations, the value of d is either 2 or 3. Let $\mathbf{\Delta} = [\delta_{ij}]$ denote the pairwise dissimilarity matrix, where δ_{ij}, $i, j = 1, 2, ..., N$ refers to the dissimilarity between the objects i and j. The resulting embedding in the d dimensional space is represented by $\mathbf{X} = [\mathbf{x}_1|\mathbf{x}_2|, ..., |\mathbf{x}_N]^T \in \mathbb{R}^{N\times d}$. That is, the i-th object is mapped to $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{id})^T \in \mathbb{R}^{d\times 1}$, where x_{ij} is the j-th coordinate of $\mathbf{x}_i$. Let $\mathbf{D}(\mathbf{X}) = [d_{ij}(\mathbf{X})] \in \mathbb{R}^{N\times N}$ denote the distance matrix having as ij-th element the ℓ_2 norm between $\mathbf{x}_i$ and $\mathbf{x}_j$, i.e., $d_{ij}(\mathbf{X}) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. It can be shown that

$$[\mathbf{D}(\mathbf{X})]^2 = \mathrm{diag}(\mathrm{diag}(\mathbf{X}\mathbf{X}^T))\,\mathbf{E} + \mathbf{E}\,\mathrm{diag}(\mathrm{diag}(\mathbf{X}\mathbf{X}^T)) - 2\mathbf{X}\mathbf{X}^T \tag{1}$$

where $\mathbf{E}$ is a $N \times N$ matrix with all its elements equal to one and $[\mathbf{D}(\mathbf{X})]^2$ denotes the Hadamard product of $\mathbf{D}(\mathbf{X})$ with itself. Accordingly, the elements of the matrix $[\mathbf{D}(\mathbf{X})]^2$ are the squared distances.

The MDS is a non-linear optimization problem, where the minimization of distance distortions is sought. A least squares loss func-

tion that measures the goodness of fit is the raw stress defined as:

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

The fragility of stress to outliers, as a least squares loss function, has inspired many researchers to investigate possible alternatives in order to eliminate the influence of gross errors. The cost function $\left\|\mathbf{\Delta}^2 - \mathbf{D}^2\right\|_1$ was employed in the robust Euclidean embedding (REE) [10] in order to minimize the influence of noise. Other related ideas can be found in [11, 12].

In the robust MDS (RMDS) [13], the variable o_{ij} is inserted to model any outlier in δ_{ij}. That is, each dissimilarity element is modeled as $\delta_{ij} = d_{ij}(\mathbf{X}) + o_{ij} + \epsilon_{ij}$, where ϵ_{ij} denotes a zero-mean independent random variable modeling the nominal errors. In addition, the ℓ_1 norm of the $N \times N$ outlier matrix is included in the optimization problem due to the sparseness of the outliers, suggesting that a small amount of them admits non-zero values. Accordingly, the following optimization problem is solved by alternating minimization [13]:

$$(\hat{\mathbf{O}}, \hat{\mathbf{X}}) = \underset{\mathbf{O},\mathbf{X}}{\operatorname{argmin}} \sum_{i<j}^{N}(\delta_{ij} - d_{ij}(\mathbf{X}) - o_{ij})^2 + \lambda_1 \sum_{i<j}^{N}|o_{ij}| \quad (3)$$

where $\|\mathbf{O}\|_1 = 2\sum_{i<j}|o_{ij}|$. The solution of (3) at iteration $t+1$ is given in closed form as [13]:

$$o_{ij}^{(t+1)} = S_{\lambda_1}(\delta_{ij} - d_{ij}(\mathbf{X}^{(t)})) \quad (4)$$

$$\mathbf{X}^{(t+1)} = \mathbf{L}^{\dagger}\,\mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\,\mathbf{X}^{(t)} \quad (5)$$

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, which is 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 (5), $\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})$ is the Laplacian matrix having elements:

$$[\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})]_{ij} = \begin{cases} -(\delta_{ij} - o_{ij})\, d_{ij}^{-1}(\mathbf{X}) & (i,j) \in \mathbb{S}(\mathbf{O}, \mathbf{X}) \\ 0 & (i,j) \in \mathbb{T}(\mathbf{O}, \mathbf{X}) \\ -\sum_{k=1, k\neq i}^{N}[\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})]_{ik} & (i,j) \in \mathbb{Q}(\mathbf{O}, \mathbf{X}) \end{cases} \quad (6)$$

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

It is seen that (5) is still vulnerable to outliers. M-estimators, which constitute a generalization of maximum likelihood estimators [14], replace the least squares loss function, which is sensitive to outliers, with another, which increases less than the squared error and thus it is less fragile to outliers. Accordingly, it is proposed to seek for the M-estimator of $\mathbf{X}$ passing the residual $\mathbf{L}\mathbf{X} - \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$ through a function $\phi(\cdot)$ that is non-negative and differentiable with respect to $\mathbf{X}$ and to impose a regularization term associated to the Frobenius norm of $\mathbf{X}$, i.e.,

$$\hat{\mathbf{X}} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ \phi(\mathbf{L}\mathbf{X} - \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}) + \lambda_2 \left\|\mathbf{X}\right\|_F^2 \right\} \quad (7)$$

The optimization problem (7) is solved by HQ minimization in Section 4 and links are established with the maximum correntropy criterion, which is related to the Welsch M-estimator, in Section 3. By doing so, a unified framework emerges that extends the work in [13].

3. CORRENTROPY

The (cross) correntropy was first introduced as a generalized correlation function [15]. It is a nonlinear similarity metric between two arbitrary random variables W and Y defined as $V_\sigma(W, Y) = E[g_\sigma(W - Y)]$, where $E[\cdot]$ is the expectation operator and $g_\sigma(x) = \exp(-\frac{x^2}{2\sigma^2})$ is the Gaussian kernel with kernel size σ [16]. When a finite amount of data (y_i, w_i), $i = 1, 2, \ldots, N$ is available, the sample estimator of correntropy is used, i.e.:

$$\hat{V}_\sigma(W, Y) = \frac{1}{N}\sum_{i=1}^{N} g_\sigma(w_i - y_i). \quad (8)$$

The correntropy measure is symmetric, positive, and bounded, attaining a maximum for $W = Y$. Its properties depend on the kernel size, whose selection is application specific. For two random vectors $\underline{W} = (w_1, w_2, \ldots, w_N)^T$ and $\underline{Y} = (y_1, y_2, \ldots, y_N)^T$, the Correntropy Induced Metric (CIM) is defined as [16]

$$CIM(\underline{W}, \underline{Y}) = \left[g_\sigma(0) - \frac{1}{N}\sum_{i=1}^{N} g_\sigma(w_i - y_i) \right]^{1/2}. \quad (9)$$

The CIM possesses the properties of symmetry, non-negativity and triangle inequality. In addition, $CIM(\underline{W}, \underline{Y}) = 0$, if and only if $\underline{W} = \underline{Y}$ [16]. The Maximum Correntropy Criterion (MCC) aims at maximizing $\hat{V}_\sigma(W, Y)$. Since the CIM is a decreasing function of correntropy, the maximization of correntropy is equivalent to the minimization of the CIM. The Gaussian kernel makes the MCC a local criterion [16], restricting the analysis to a local region of the joint space of w and y. Indeed, the correntropy is determined by the kernel function along the line $w = y$. On the contrary, the mean squared error (MSE) is a global criterion, where all the sample errors conduce considerably to its estimation. For gross errors, the MSE increases quadratically, while the CIM is close to 1, mitigating the effect of outliers.

The correntropy is closely related to the M-estimators [16]. By setting $\phi(x) = 1 - g_\sigma(x)$, the CIM is equivalent to the Welsch M-estimator. The MCC has proven to be an appropriate similarity metric in non-linear, non-Gaussian, signal processing applications, such as robust regression [16], pattern recognition [17], feature selection [18], and subspace clustering [19, 20].

4. AN HQ FRAMEWORK FOR MDS WITH OUTLIERS

In this section, the optimization problem (7) is solved with HQ minimization [21, 22]. There are two forms of the HQ, namely the additive form and the multiplicative one. Here, due to space limitations, we are confined to the latter. Let $\phi(x)$ be a potential function that satisfies the conditions in [21]. Then, a conjugate (dual) function $\psi(\cdot)$ exists for fixed x, such that [21]:

$$\phi(x) = \inf_{p \in \mathbb{R}} \{Q(x, p) + \psi(p)\} \quad (10)$$

where $Q(x, p) = px^2$ is a quadratic function of x, and p is an auxiliary variable determined by the minimizer function $\delta(\cdot)$ related to $\phi(\cdot)$. Table 1 lists the potential function $\phi(x) : \mathbb{R} \to \mathbb{R}$ and the minimizer function $\delta(x) : \mathbb{R} \to \mathbb{R}$ for the multiplicative form of the HQ for various M-estimators.

Let $\mathbf{Y} = \mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$. The objective function in (7) takes the form:

$$J(\mathbf{X}, \mathbf{p}) = \sum_{i=1}^{N} p_i \left\|(\mathbf{L}\mathbf{X} - \mathbf{Y})^i\right\|_2^2 + \sum_{i=1}^{N}\psi(p_i) + \lambda_2 \left\|\mathbf{X}\right\|_F^2 \quad (11)$$

Table 1: Potential functions of M-estimators and their minimizer functions for the multiplicative form of HQ

M-estimator	*Potential Function*	*Minimizer Function*
ℓ_2	$\phi(x) = x^2/2$	$\delta(x) = 1$
ℓ_p	$\phi(x) = \frac{\lvert x\rvert^p}{p}\ p \in (1,2]$	$\delta(x) = \lvert x\rvert^{p-2}$
Fair	$\phi(x) = a^2(\frac{\lvert x\rvert}{a} - \log(1+\frac{\lvert x\rvert}{a}))$	$\delta(x) = \frac{1}{1+\frac{\lvert x\rvert}{a}}$
Welsch	$\phi(x) = \frac{a^2}{2}(1-\exp(-\frac{x^2}{a^2}))$	$\delta(x) = \exp(-\frac{x^2}{a^2})$
Cauchy	$\phi(x) = \frac{a^2}{2}\log(1+(\frac{x}{a})^2)$	$\delta(x) = \frac{1}{1+(\frac{x}{a})^2}$

where $\mathbf{p}$ is the vector of the auxiliary variables. It is seen that (11) depends on the weighted sum of the squared ℓ_2 norms of the residuals $\mathbf{LX} - \mathbf{Y}$ for each row. Let $(\hat{\mathbf{X}}, \hat{\mathbf{p}}) = \underset{\mathbf{X},\mathbf{p}}{\operatorname{argmin}} J(\mathbf{X}, \mathbf{p})$. Due to the fact that the auxiliary variables depend only on the minimizer function $\delta(\cdot)$, the terms $\psi(\cdot)$ can be omitted as being fixed, when we minimize w.r.t. $\mathbf{X}$. Thus, a local minimizer $(\mathbf{X}, \mathbf{p})$ is estimated using the following alternating minimization:

$$p_i^{(t+1)} = \delta\left(\left\|(\mathbf{LX}^{(t)} - \mathbf{Y})^i\right\|_2\right) \tag{12}$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}}\Big\{\operatorname{tr}((\mathbf{LX}-\mathbf{Y})^T\mathbf{P}^{(t+1)}(\mathbf{LX}-\mathbf{Y})) + \lambda_2\operatorname{tr}(\mathbf{X}^T\mathbf{X})\Big\} \tag{13}$$

where $\mathbf{P}^{(t+1)} = \operatorname{diag}(\mathbf{p}^{(t+1)})$ is a diagonal matrix with ii-th element equal to $p_i^{(t+1)}$. Setting the derivative of (13) w.r.t. $\mathbf{X}$ equal to zero, a closed-form solution is obtained, i.e.:

$$\mathbf{X}^{(t+1)} = (\mathbf{L}^T\,\mathbf{P}^{(t+1)}\,\mathbf{L} + \lambda_2\mathbf{I})^{-1}\,\mathbf{L}^T\,\mathbf{P}^{(t+1)}\,\mathbf{Y}. \tag{14}$$

At each iteration, the auxiliary variables $p_i^{(t+1)}$ represent the weight that regulates the impact of $\left\|(\mathbf{LX}-\mathbf{Y})^i\right\|_2$. The introduction of M-estimators reduces the influence of the outliers, since $p_i^{(t+1)}$ always admits a low weight, as is manifested by the presence of $\delta(\cdot)$ in (12) that is associated to the potential function ϕ of an M-estimator. The multiplicative form of the HQ optimization is essentially an iterative reweighted least-squares minimization, that has been used in robust regression in order to mitigate the outliers influence. The complete procedure for the solution of (7) by the multiplicative form of HQ is outlined in Algorithm 1. The initial configuration $\mathbf{X}^{(0)}$ is chosen randomly, while the initial outlier matrix $\mathbf{O}^{(0)}$ is set to zero. The basic property $J(\mathbf{X}^{(t+1)}, \mathbf{p}^{(t+1)}) \leq J(\mathbf{X}^{(t)}, \mathbf{p}^{(t+1)}) \leq J(\mathbf{X}^{(t)}, \mathbf{p}^{(t)})$ of the HQ guarantees that the objective function is reduced at each iteration until its convergence [21].

Algorithm 1 Multiplicative form of the HQ minimization for MDS (MHQMDS)

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 (4)
3: Update $p_i^{(t+1)}$ via (12) with $\mathbf{L}_+$ as in (6)
4: Update $\mathbf{X}^{(t+1)}$ via (14)
5: **end for**

5. NUMERICAL TESTS

The multiplicative form of the HQ minimization for the MDS (MHQMDS) was implemented in Matlab and tested on several sets of dissimilarity matrices $\mathbf{\Delta}$. In order to evaluate and benchmark the MHQMDS, three well known MDS techniques were implemented in the same environment and tested on the same dissimilarity matrices. These techniques were: a) the popular SMACOF algorithm [8], b) the subgradient version of the REE algorithm [10], and c) the RMDS algorithm [13]. In all techniques, the authors' recommendations were strictly followed, while the implementation was intended to achieve the best possible performance.

The embedding quality for each algorithm was evaluated with respect to four figures of merit: a) the normalized outlier-free stress $\sigma(\hat{O}, \hat{X}) = \sqrt{\frac{\sum_{(i,j)\in Q}(\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{(i,j)\in Q}\delta_{ij}^2}}$, as in [13], where Q denotes the set of outlier-free dissimilarities (i.e., when $[\mathbf{O}]_{ij} = 0$), b) the number of outliers $\hat{S}$ as in [13], c) the raw stress $\sigma_r(\mathbf{X})$ between the distances of the final embedding and the initial dissimilarities, and d) the Procrustean goodness-of-fit ϱ, standardized by a measure of the scale for $\mathbf{X}$[1]. The last criterion can be used only for fixed configurations.

In order to assess the implemented methods, 100 Monte Carlo simulations were run using a different random initial configuration $\mathbf{X}^{(0)}$ on each run. The reported figures of merit refer to the run, where the RMDS algorithm has exhibited the minimum value in raw stress $\sigma_r(\mathbf{X})$, namely when the final embedding was closer to the initial configuration. The algorithms RMDS and MHQMDS terminated when $\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F / \left\|\mathbf{X}^{(t+1)}\right\|_F$ was less than 10^{-6} or when the number of iterations reached 5000.

The data set, that was used to evaluate the performance of the MHQMDS algorithm, comprised a rectangular with $N = 100$ points in the two-dimensional space. The bottom-left point was at $(1, 1)$, while all points were equidistant from their vertical and horizontal neighbors by one unit. Each element of the initial dissimilarity matrix $\mathbf{\Delta}$ was contaminated with a background error ϵ_{ij}, derived from a zero mean truncated Gaussian distribution with variance $\sigma^2 = 0.1$ and threshold $-d_{ij}(\mathbf{X})$, in order to avert negative values in $\mathbf{\Delta}$. The indices of the outliers were chosen randomly, while their values were derived from a uniform distribution in [0,40]. The outlier contamination percentage ϖ was set at 40%. Let a_h be the parameter of the Huber M-estimator. Taking into account the equivalence with Huber M-estimator for $\lambda_1 = 2a_h$ [23] and that $a_h = 1.345\sigma$ yields 95% asymptotic efficiency for the normal distribution [24], λ_1 was set to 0.851 for both the RMDS and the MHQMDS algorithms.

Table 2 gathers the figures of merit related to the embedding quality delivered by the SMACOF, the REE and the RMDS algorithms. Due to the lack of space, only the raw stress $\sigma_r(\mathbf{X})$ of the MHQMDS algorithm is plotted for various values of $\lambda_2 \in [1, 100]$ in Figure 1. The plots of the normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$, the number of outliers $\hat{S}$ and the standardized Procrustean goodness-of-fit ϱ are roughly the same with that of the raw stress. The parameter a was set equal to 316.228, 14, and 10 for the Welsch, the Cauchy and the Fair M-estimator, respectively. It is obvious that these M-

[1]In Matlab, the measure of the scale for X is given by `sum(sum((X-repmat(mean(X,1),size(X,1),1)).^2,1))`.

Table 2: Figures of merit judging the embedding quality obtained by the SMACOF, the REE, and the RMDS algorithms applied to the square data set whose 40% of elements have been corrupted by outliers.

$\varpi = 40\%$	SMACOF	REE	RMDS
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.6720	0.7892	0.0452
Estimated outliers $\hat{S}$	-	-	3329
Procrustean goodness-of-fit ϱ	0.9029	0.0113	0.0063
Raw Stress $\sigma_r(\mathbf{X})$	395494	2010.1	1730.9

estimators, when they are employed in the multiplicative form of the HQ, outperform the state of the art approaches for a wide range of values admitted by λ_2. The performance of ℓ_p M-estimator for $p = 1.999$ was comparable to that of the Welsch M-estimator.

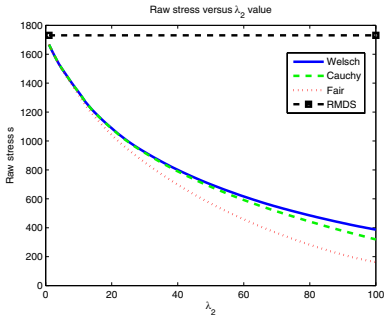

Fig. 1: Raw stress $\sigma_r(\mathbf{X})$ of the MHQMDS algorithm.

The embeddings delivered by the RMDS and the MHQMDS algorithms, after being matched to the true embedding via Procrustes' analysis, are demonstrated in Figure 2. The MHQMDS embedding was obtained by the Welsch M-estimator for $\lambda_2 = 100$. It is obvious that the MHQMDS embedding, whose raw stress $\sigma_r(\mathbf{X})$ is equal to 386.7, is slightly closer to the true embedding than that of the RMDS, whose $\sigma_r(\mathbf{X})$ is 1730.9. This is also validated by the values of the Procrustean goodness-of-fit ϱ, which were 0.0063 and 0.0019 for the RMDS and MHQMDS algorithms, respectively.

5.1. Discussion

It is apparent that the proposed algorithm outperforms the state of the art approaches, since it accomplishes a more accurate embedding than the RMDS for a wide range of λ_2 values. The SMACOF is extremely inefficient, while the REE obtains a better embedding than the SMACOF, but still this is inferior than that of the RMDS.

The efficiency of M-estimators is determined highly by the proper selection of the parameter a, related to the kernel size of the Welsch potential function. The experimental results validate that when a is large, the region where the ℓ_2 norm is applicable (and consequently the MSE applies) expands. Under these circumstances, the performance of the Welsch M-estimator approximates that of the ℓ_2 M-estimator. Contrarily, a small kernel size shrinks the region where the ℓ_2 norm applies (while the ℓ_1 and ℓ_0 regions are enlarged). In such a case, the range of values of λ_2 for which the MHQMDS is more efficient than the RMDS, w.r.t. raw stress, is much smaller. However, this choice leads to a smaller λ_2 value, where the raw

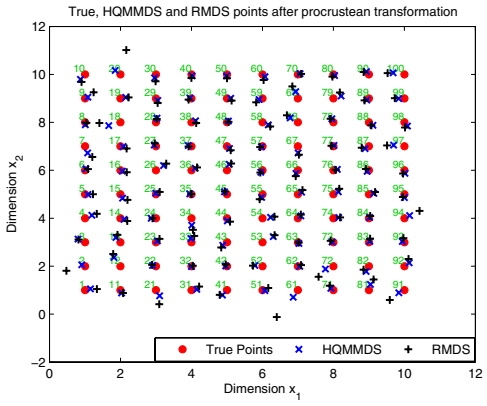

Fig. 2: MHQMDS and RMDS embeddings after being matched to the true embedding via Procrustes' analysis.

stress $\sigma_r(\mathbf{X})$ attains its minimum, accelerating significantly the finding of the optimal embedding. These remarks were also validated for the Cauchy and Fair M-estimators. However, the value of a that induces an equivalent performance with the ℓ_2 M-estimator is different for each M-estimator and is highly data-dependent.

Parameter setting depends on the user's objective. If the objective is a large range of values for λ_2 where the MHQMDS is more efficient than the RMDS w.r.t. the raw stress, then a large value of a should be chosen. On the contrary, if the objective is a fast finding of the true configuration, then a small value of a is sufficient.

If the initial dissimilarity matrix is not available, the normalized outlier free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ and the number of outliers $\hat{S}$ can only be used as figures of merit. Under these circumstances, the MHQMDS algorithm is implemented for a plausible range of values for λ_2, selecting a small value of the parameter λ_1, and then the embedding with the minimum value of the $\hat{S}$ is selected. Extensive experimental results have proven that this embedding is close to the embedding where the raw stress $\sigma_r(\mathbf{X})$ is minimum, which indicates that the true configuration is well approximated.

In many cases, the multiplicative form of the HQ minimization for the MDS entailed fewer iterations in order to converge than the RMDS, rendering it slightly faster. It can be proven that the computational complexity of the MHQMDS algorithm, which involves alternating updates of $\mathbf{O}$, $\mathbf{p}$ and $\mathbf{X}$, is $O(N^3)$ at each iteration.

6. CONCLUSIONS

A new efficient HQ framework, using the multiplicative form, has been introduced for solving the MDS problem in an environment contaminated by outliers. The experimental findings have demonstrated that the proposed algorithm performs substantially better than the state-of-the-art. For any given configuration contaminated with outliers, it is possible to find an M-estimator so that the proposed MHQMDS outperforms the state-of-the-art MDS approaches. Future research will address techniques for estimating the kernel size of the potential function within the MHQMDS.

Acknowledgments. This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operation Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS - UOA - ERASITECHNIS MIS 375435.

7. REFERENCES

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

[2] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000.

[3] E. R. Gansner, Y. Koren, and S. C. North, "Graph drawing by stress majorization," in *Proc. 12th Int. Conf. Graph Drawing (GD'04)*, Jnos Pach, Ed., Berlin, 2005, vol. LNCS 3383, pp. 239–250, Springer-Verlag.

[4] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen, "Data visualization with multidimensional scaling," *Journal of Computational and Graphical Statistics*, vol. 17, no. 2, pp. 444–472, 2008.

[5] G. Zigelman, R. Kimmel, and N. Kiryati, "Texture mapping using surface flattening via multidimensional scaling," *IEEE Trans. Visualization and Computer Graphics*, vol. 8, no. 2, pp. 198–207, 2002.

[6] A. Pal, "Localization algorithms in wireless sensor networks: Current approaches and future challenges," *Network Protocols and Algorithms*, vol. 2, no. 1, pp. 45–74, 2010.

[7] E. Cambria, Y. Song, H. Wang, and N. Howard, "Semantic multi-dimensional scaling for open-domain sentiment analysis," *IEEE Intelligent Systems*, vol. 99, no. 2, pp. 44–51, 2014.

[8] 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.

[9] I. Spence and S. Lewandowsky, "Robust multidimensional scaling," *Psychometrika*, vol. 54, no. 3, pp. 501–513, 1989.

[10] L. Cayton and S. Dasgupta, "Robust Euclidean embedding," in *Proc. 23rd Int. Conf. Machine Learning*, June 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, June 1987, pp. 455–462.

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

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

[14] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 55, pp. 73–101, 1964.

[15] I. Santamara, P. P. Pokharel, and J. C. Principe, "Generalized correlation function: definition, properties, and application to blind equalization," *IEEE Trans. Signal Processing*, vol. 54, no. 6, pp. 2187–2197, June 2006.

[16] 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.

[17] R. He, W. S. Zheng, B. G. Hu, and X. W. Kong, "A regularized correntropy framework for robust pattern recognition," *Neural Computation*, vol. 23, no. 8, pp. 2074–2100, August 2011.

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

[19] 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*, December 2013, pp. 1801–1808.

[20] Y. Zhang, Z. Sun, R. He, and T. Tan, "Robust subspace clustering via half-quadratic minimization," in *Proc. IEEE Int. Conf. Computer Vision*, December 2013, pp. 3096–3103.

[21] 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, Oct 2005.

[22] R. He, W. S. Zheng, T. Tan, and Z. Sun, "Half-quadratic-based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 261–275, 2014.

[23] 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.

[24] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image and Vision Computing*, vol. 15, pp. 59–76, 1997.