CORRENTROPY BASED ROBUST MULTIDIMENSIONAL SCALING APPLIED TO FACES

Fotios Mandanas, Constantine Kotropoulos[*]

Aristotle University of Thessaloniki
Department of Informatics
Thessaloniki 54124, GREECE
fmandan@gmail.com, costas@aiia.csd.auth.gr

ABSTRACT

Here, we are interested in obtaining a two-dimensional embedding of face-pose images that preserves their local structure captured by the pair-wise distances among them by using multidimensional scaling (MDS). The MDS problem is formulated as maximization of a correntropy criterion, which is solved by half-quadratic optimization in a multiplicative formulation. By doing so, the MDS copes with an initial dissimilarity matrix contaminated with outliers, because the correntropy criterion is closely related to the Welsch M-estimator. The proposed algorithm is coined as Multiplicative Half-Quadratic MDS (MHQMDS). Its performance is assessed for potential functions associated to various M-estimators have been tested. 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, outperforms the aforementioned state-of-the-art competing techniques.

Index Terms— Multidimensional scaling, robustness, M-estimators, correntropy, half-quadratic optimization, face-pose images

1. INTRODUCTION

Multidimensional Scaling (MDS) 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]. Since then, 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]. MDS has found many applications in the forensics and biometrics. For example, it was applied to Earth Mover's Distance (EMD) data calculated between 60 different writers in order to visualize each writer's feature in population, assisting forensic handwriting experts in the process of writer verification by visualizing the diversity of overall shapes of digit handwritings [7]. The MDS has also been used to visualize the similarities between speaker models or between models and data vectors from recordings in two dimensions [8]. Recently, a semantic MDS for open-domain sentiment analysis was developed [9]. In addition to the aforementioned applications that deal with the so-called metric MDS, the MDS was also applied to non-metric (or ordinal) data [10–12]. The latter approach has also proved useful within forensics, such as forensic phychology research [13] or DNA sequence analysis [14]. In the following, we shall confined ourselves to the metric MDS.

Subspace clustering or hybrid linear modeling [15, 16] major premise is that the total variance of the data in the aforementioned tasks is contained in a small number of principal axes. Even if the measured data are high-dimensional, their intrinsic dimensionality is usually much lower. Let face images be represented as vectors by using lexicographical ordering. Although face images lie on an input space of high dimensionality (i.e., $d = 4096$ for images of size 64×64), their meaningful structure exhibits much fewer independent degrees of freedom. For example, it has been demonstrated that face images actually lie on a three-dimensional manifold, which is parameterized by the two pose variables (left-right pose, up-down pose) and one lighting direction variable [2]. Here, we are addressing the following problem: Given a dissimilarity matrix among face images whose elements are corrupted by nominal errors as well as outliers how one may accurately visualize these images in two dimensions by finding a two-dimensional embedding of face-poses.

The traditional algorithms for the solution of the MDS problem, like the classical MDS [1] and the scaling by majorizing a complicated function (SMACOF) [17], 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 [18, 19], because the noise is propagated to each element of the distance matrix through the double-centering process involved (cf. next section).

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. Particularly, when the dissimilarity matrix is contaminated with outliers the following novel contributions are made: 1) A framework, that is based on half-quadratic (HQ) minimization in combination with M-estimators, is developed in order to estimate the MDS embedding; 2) An efficient algorithm for finding the MDS solution is proposed, 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 (i, j) 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 $\mathrm{diag}(\mathbf{X})$

[*]Supported by the COST Action IC 1106 "Integrating Biometrics and Forensics for the Digital Age".

978-1-4799-8105-2/15/$31.00

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 $\|\mathbf{X}\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m x_{ij}^2}$.

The remainder of this paper is structured as follows: Section 2 is devoted to MDS approaches that reduce the influence of outliers. Special emphasis is given to the Robust MDS (RMDS), proposed in [20]. An overview of the correntropy measure is presented in Section 3. The proposed unified framework, where the MDS is treated as maximization of a correntropy criterion, is presented in Section 4. Experimental results for face-pose visualization are demonstrated in Section 5 and discussed in Section 6. Finally, Section 7 concludes the paper and indicates future research directions.

2. ROBUST MDS APPROACHES

Let N denote the number of objects and d be the embedding dimension. For visualization purposes, d admits the value 2 or 3. Let $\mathbf{\Delta} = [\delta_{ij}]$ denote the pairwise dissimilarity matrix, where δ_{ij}, $i, j = 1, 2, \ldots, N$ 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

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

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.

The MDS yields a non-linear optimization problem, where the minimization of distance distortions is sought. A least squares (LS) loss function that represents the goodness of fit between δ_{ij} and $d_{ij}(\mathbf{X})$ is the raw stress defined as:

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

Being a LS loss function, the raw stress is fragile to outliers. This fragility has inspired many researchers to investigate possible alternatives in order to eliminate the influence of gross errors. For example, the cost function $\left\|\mathbf{\Delta}^2 - \mathbf{D}^2\right\|_1$ was employed in the Robust Euclidean Embedding (REE) [19] in order to minimize the influence of noise. Other related ideas can be found in [21, 22].

In the RMDS [20], 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. Moreover, due to the sparseness of the outliers, the ℓ_1 norm of the $N \times N$ outlier matrix is included in the optimization criterion, suggesting that a small amount of them admits non-zero values. Accordingly, the following optimization problem is solved by alternating minimization [20]:

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

where $\|\mathbf{O}\|_1 = 2\sum_{i<j}|o_{ij}|$. The solution of (3) at iteration $t+1$ is given in closed form as [20]:

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

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

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:

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

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 $\mathbf{O}^{(0)}$.

(5) is still vulnerable to outliers, because of the sensitivity of the LS loss function to outliers. If M-estimators (i.e., a generalization of Maximum Likelihood Estimators [23]) replace the LS loss function with another, which increases less than the squared error, the resulting objective function will be 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.,

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

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 [20].

3. CORRENTROPY

The (cross) correntropy was first introduced as a generalized correlation function [24]. 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 $g_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x^2}{2\sigma^2})$ is the Gaussian kernel with kernel size σ [25]. 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.:

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

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 $\mathbf{W} = (w_1, w_2, \ldots, w_N)^T$ and $\mathbf{Y} = (y_1, y_2, \ldots, y_N)^T$, the

Correntropy Induced Metric (CIM) is defined as [25]

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

The CIM possesses the properties of symmetry, non-negativity and triangle inequality. In addition, $CIM(\mathbf{W},\mathbf{Y}) = 0$, if and only if $\mathbf{W} = \mathbf{Y}$ [25]. 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 [25], 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 [25]. By setting $\phi(x) = 1 - g_\sigma(x)$, the CIM is equivalent to the Welsch M-estimator. The MCC as a similarity metric has proven to be appropriate in non-linear, non Gaussian signal processing applications, such as robust regression [25], pattern recognition [26], feature selection [27], and subspace clustering [28, 29].

4. AN HQ FRAMEWORK FOR MDS WITH OUTLIERS

In this section, the optimization problem (7) is solved with HQ minimization [30, 31]. There are two forms of the HQ, 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 [30]. Then for fixed x, a conjugate (dual) function $\psi(\cdot)$ exists, such that [30]:

$$\phi(x) = \inf_{p\in\mathbb{R}}\{\frac{1}{2}px^2 + \psi(p)\} \quad (10)$$

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

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	$\phi(x) = \frac{a^2}{2}\log(1 + (\frac{x}{a})^2)$	$\delta(x) = \frac{1}{1+(\frac{x}{a})^2}$

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

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

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 $\mathbf{L}\mathbf{X} - \mathbf{Y}$ 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 $\psi(\cdot)$ are fixed and can be omitted, when we minimize w.r.t. $\mathbf{X}$. Thus, a local minimizer $(\mathbf{X},\mathbf{p})$ is estimated using the following alternating minimization:

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

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

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.:

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

At each iteration, the auxiliary variables $p_i^{(t+1)}$ provide the weight that regulates the impact of $\left\|(\mathbf{L}\mathbf{X} - \mathbf{Y})^i\right\|_2$. 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 $J(\mathbf{X}^{(t+1)},\mathbf{p}^{(t+1)}) \le J(\mathbf{X}^{(t)},\mathbf{p}^{(t+1)}) \le J(\mathbf{X}^{(t)},\mathbf{p}^{(t)})$ of the HQ guarantees that the objective function is reduced at each iteration until its convergence [30].

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. EXPERIMENTAL RESULTS

The data set used to evaluate the performance of the MHQMDS algorithm comprises a subset of $N = 100$ from a total of 698 face images of size 64×64 with different poses and lighting directions [2]. Although these images lie on a high dimensionality input space (i.e., $d = 4096 = 64 \times 64$), their intrinsic structure exhibits fewer independent degrees of freedom. More specifically, the images lie on a three-dimensional manifold, which can be parameterized by two pose variables (left-right pose, up-down pose) and one lighting direction variable [2].

Trying simply to obtain a two-dimensional embedding, preserving the local structure captured by the pair-wise dissimilarities

among the face images, the initial 100×100 dissimilarity matrix $\mathbf{\Delta}$ was first computed. Let the i-th face-pose image, $i = 1, 2, \ldots, 100$ be represented by $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ij}, \ldots, y_{i,4096})^T \in \mathbb{R}^{4096\times 1}$, where y_{ij} is the j-th coordinate of $\mathbf{y}_i$, with $j = 1, 2, \ldots, 4096$. The ij-th element of the matrix $\mathbf{\Delta}$ refers to the dissimilarity between the face images i and j, namely the ℓ_2 norm between $\mathbf{y}_i$ and $\mathbf{y}_j$. The distinct pairwise dissimilarities of this matrix are $\frac{N(N-1)}{2} = 4950$. The dissimilarity matrix $\mathbf{\Delta}$ was artificially contaminated by $\varpi = 500/4950 = 10.101\%$ outliers, which were drawn from a uniform distribution in $[0, 3\max\{\delta_{ij}\}]$. The indices of the outliers were chosen randomly. The parameter λ_1 in the RMDS algorithm was set to 24.15 in order to identify $\hat{S} = 500$ outliers. The same value of λ_1 was used in the MHQMDS algorithm.

In order to evaluate and benchmark the MHQMDS, three well known MDS techniques were implemented in the same environment (Matlab) and tested on the same dissimilarity matrix. These techniques were: a) the popular SMACOF algorithm [17], b) the subgradient version of the REE algorithm [19], and c) the RMDS algorithm [20]. 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 three figures of merit: a) the normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}}) = \sqrt{\frac{\sum_{(i,j)\in Q}(\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{(i,j)\in Q}\delta_{ij}^2}}$, as in [20], 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 [20], c) the raw stress $\sigma_r(\hat{\mathbf{X}})$ between the distances of the final embedding and the initial non-contaminated (clean) dissimilarities. For fixed configurations, the Procrustean goodness-of-fit ϱ can also be used as a figure of merit, standardized by a measure of the scale for $\mathbf{X}$[1].

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 $\sigma_r(\hat{\mathbf{X}})$, 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.

Table 2: Figures of merit judging the embedding quality obtained by the SMACOF, the REE, and the RMDS algorithms applied to the face images subset whose 10.101% of elements have been corrupted by outliers.

$\varpi = 40\%$	SMACOF	REE	RMDS
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.5854	0.6796	0.2544
Estimated outliers $\hat{S}$	-	-	500
Raw Stress $\sigma_r(\hat{\mathbf{X}})$	$4.22\,10^5$	$3.38\,10^5$	$1.44\,10^5$

Table 2 gathers the figures of merit related to the embedding quality delivered by the SMACOF, the REE and the RMDS algorithms. The normalized outlier free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ and the raw stress $\sigma_r(\hat{\mathbf{X}})$, after the implementation of the SMACOF algorithm on the non-contaminated (clean) data, were equal to 0.2515 and $1.3088\,10^5$, respectively. The figures of merit of the MHQMDS algorithm, for various values of $\lambda_2 \in [1, 100]$, are plotted in Figures 1 and 2. The plot of the normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ is roughly the same with that of the raw stress $\sigma_r(\hat{\mathbf{X}})$. The parameter a was set equal to 1000, 80, and 20 for the Welsch, the Cauchy and the Fair M-estimator, respectively.

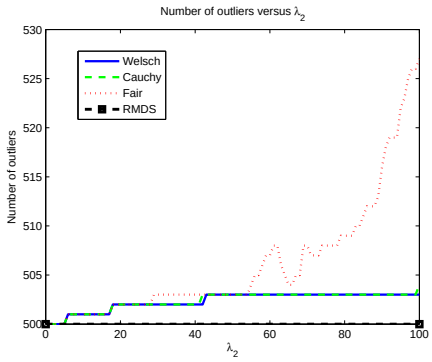

Fig. 1: Estimated number of outliers $\hat{S}$ of the MHQMDS algorithm.

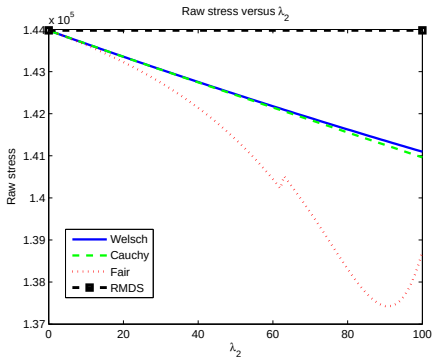

Fig. 2: Raw stress $\sigma_r(\hat{\mathbf{X}})$ of the MHQMDS algorithm.

It is obvious that these M-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 the regularization parameter λ_2. The normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ and the raw stress $\sigma_r(\hat{\mathbf{X}})$ for the Welsch, Cauchy and Fair M-estimators are less than the corresponding values of the RMDS algorithm for $\lambda_2 \in [1, 100]$. The performance of ℓ_p M-estimator for $p = 1.999$ was comparable to that of the Welsch M-estimator.

The embeddings delivered by the SMACOF and MHQMDS algorithms are demonstrated in Figure 3. The MHQMDS embedding was obtained by the Welsch M-estimator for $\lambda_2 = 100$. In this case, the raw stress $\sigma_r(\hat{\mathbf{X}})$ of the MHQMDS algorithm was equal to $1.41\,10^5$. The SMACOF embedding was being matched to that of the MHQMDS algorithm via Procrustes' analysis. It is seen that there is a great differentiation, which is also validated by the large value of the standardized Procrustean goodness-of-fit ϱ_1, which is 0.4027. The same conclusions can be drawn by the REE embedding, whose ϱ_1 is equal to 0.8296. The RMDS and MHQMDS embeddings are approximately the same, as illustrated in Figure 4, even though the MHQMDS algorithm exhibits lower raw stress $\sigma_r(\hat{\mathbf{X}})$ and normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ than the RMDS.

The SMACOF embedding of the face-pose images subset on the clean data, where a sample of the real input images is superimposed, is depicted in Figure 5. The SMACOF embedding on the clean data

[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))`.

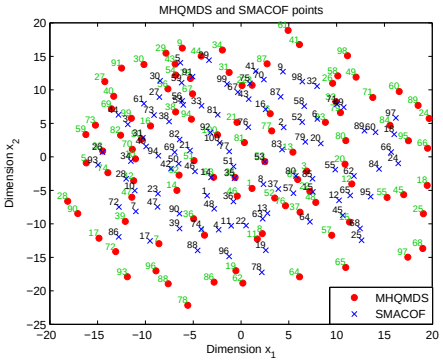

Fig. 3: SMACOF and MHQMDS embeddings on the face-pose images subset whose 10.101% of elements have been corrupted by outliers.

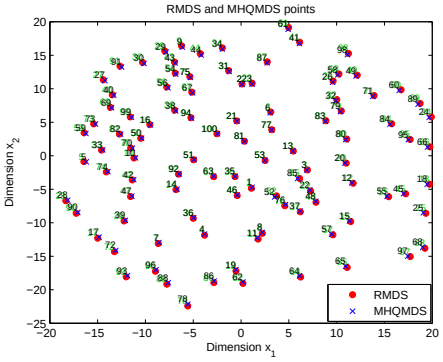

Fig. 4: RMDS and MHQMDS embeddings on the face-pose images subset whose 10.101% of elements have been corrupted by outliers.

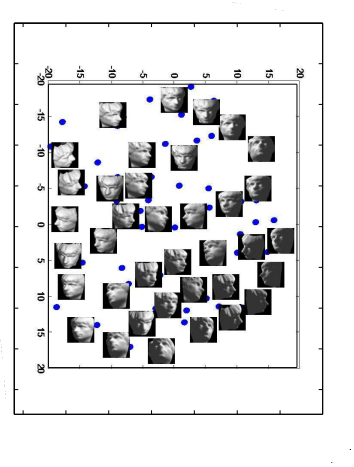

Fig. 5: SMACOF embedding on the clean subset of face-pose images with the real images superimposed on the embedding.

and the MHQMDS embedding on the corrupted data are contrasted in Figure 6. The latter was matched to the SMACOF embedding via Procrustes' analysis. It is apparent that the proposed algorithm MHQMDS on the corrupted subset preserves the embedding structure obtained by the SMACOF algorithm on the non-contaminated subset. This is also validated by the low value of the standardized Procrustean goodness-of-fit ϱ_1, which is 0.0149. The corresponding value of the RMDS algorithm is 0.0155, which demonstrates again the better performance of the MHQMDS than that of the RMDS.

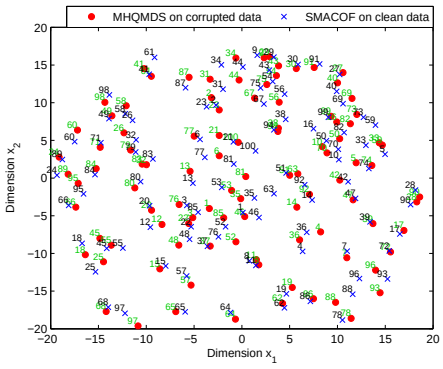

Fig. 6: Embeddings provided by the SMACOF on the clean data and the MHQMDS on the corrupted data.

6. DISCUSSION

The efficiency of the MHQMDS algorithm relies heavily on the kernel size a of the Welsch potential function. When a is large, the performance of the Welsch M-estimator approximates that of the ℓ_2 M-estimator, while a choice of a small kernel size a leads to a smaller λ_2 value, where the raw stress $\sigma_r(\mathbf{X})$ attains its minimum, accelerating the finding of the optimal approximation of the true configuration. It is worth noting that if a smaller value of the parameter λ_1 was used (i.e., $\lambda_1 = 10$ instead of 24.15), the MHQMDS would still deliver the top performance. In such a case, the plots of $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$, $\sigma_r(\hat{\mathbf{X}})$ and $\hat{S}$ would roughly be the same and would approximate the curves illustrated in Figure 2. The computational complexity of the MHQMDS algorithm, which involves alternating updates of $\mathbf{O}$, $\mathbf{p}$ and $\mathbf{X}$, is proven to be $O(N^3)$ at each iteration.

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 MHQMDS algorithm obtains its minimum raw stress $\sigma_r(\hat{\mathbf{X}})$, which indicates that the true configuration is approximated to a great extent. It has also been proven that, by selecting a small value of the parameter λ_1, the merit $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ may not provide valuable information about the embedding quality, since a higher value of this metric may correspond to a smaller value of $\sigma_r(\hat{\mathbf{X}})$.

7. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

A new efficient HQ framework, using the multiplicative form, has been introduced for solving the MDS problem in an environment

contaminated by outliers. In two-dimensional embedding of face-poses, 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 handling missing data and estimating the kernel size of the potential function within the MHQMDS. For the latter problem, one may use $a = \sqrt{\frac{\|\mathbf{LX}-\mathbf{Y}\|_F^2}{2Nd}}$ at each iteration [28].

8. 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] Y. Akao, A. Yamamoto, and Y. Higashikawal, "Assisting forensic writer verification by visualizing diversity of digit handwritings-an approach by multidimensional scaling of earth movers distance," in *Proc. 14th Int. Conf. Frontiers in Handwriting Recognition*, 2014, pp. 110–115.

[8] K. A. Weiand, J. S. Bouten, and C. J. Veenman, "Similarity visualization for the grouping of forensic speech recordings," in *Proceedings of the 2nd Int. Workshop Computational Forensics*, Berlin, Heidelberg, 2008, IWCF '08, pp. 169–180, Springer-Verlag.

[9] 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.

[10] R. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. I," *Psychometrika*, vol. 27, no. 2, pp. 125–140, June 1962.

[11] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.

[12] L. Guttman, "A general nonmetric technique for finding the smallest coordinate space for a configuration of points," *Psychometrika*, vol. 33, no. 4, pp. 469–506, December 1968.

[13] D. O' Neill and S. Hammond, "Drawing out the meaning in data: multidimensional scaling within forensic psychology research," in *The Cambridge Handbook of Forensic Psychology*, J. M. Brown and E. A. Campbell, Eds., pp. 803–812. Cambridge University Press, Cambridge, MA, 2010.

[14] E. J. Lenz and D. R. Foran, "Bacterial profiling of soil using genus-specific markers and multidimensional scaling," *Journal of Forensic Sciences*, 2010.

[15] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 52–68, March 2011.

[16] T. Zhang, A. Szlam, Y. Wang, and G. Lerman, "Hybrid linear modeling via local best-fit flats," *Int. Journal Computer Vision*, vol. 100, no. 3, pp. 217–240, 2012.

[17] 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.

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

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

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

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.