Robust Multidimensional Scaling using a Maximum Correntropy Criterion

Fotios D. Mandanas, Constantine L. Kotropoulos, *Senior Member, IEEE*

Abstract

Multidimensional Scaling (MDS) refers to a class of dimensionality reduction techniques, which represent entities as points in a low dimensional space so that the interpoint distances approximate the initial pairwise dissimilarities between entities as closely as possible. The traditional methods for solving MDS are susceptible to outliers. 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 either multiplicative or additive forms. By doing so, MDS can cope with an initial dissimilarity matrix contaminated with outliers, because the correntropy criterion is closely related to M-estimators. Three novel algorithms are derived. Their performance is assessed experimentally against three state-of-the-art MDS techniques, namely the Scaling by Majorizing a Complicated Function, the Robust Euclidean Embedding, and the Robust MDS under the same conditions. The experimental results indicate that the proposed algorithms perform substantially better than the aforementioned competing techniques.

Index Terms

Multidimensional Scaling, majorization, M-estimators, correntropy, Half-Quadratic minimization.

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Fotios Mandanas and Constantine Kotropoulos are with the Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 541 24, GREECE;

Corresponding author: F. D. Mandanas, e-mail: fmandan@gmail.com.

I. Introduction

Multidimensional Scaling (MDS) has been widely used to visualize the hidden structures among entities in a geometric space. Being a dimensionality reduction technique, MDS can be treated as a transformation yielding a geometric model (or configuration), so that the resulting interpoint distances between the entities in the new space approximate the initial pairwise dissimilarities as closely as possible. MDS seeks a configuration that corresponds to a given dissimilarity matrix, which captures the pairwise dissimilarities between the entities. MDS was inaugurated in psychology [1], [2], [3], [4]. Its spectrum of applications includes dimensionality reduction [5], graph drawing [6], [7], texture mapping on arbitrary surfaces [8], and localizing nodes in a wireless sensor network [9] to mention a few.

This paper extends the preliminary results presented in [10]. It is inspired by the work in [11]. It is motivated by the fact that when outliers are present, the use of M-estimators in the algorithms solving the MDS problem mitigates their effect more efficiently than the state of the art. In summary, the contributions of the paper are: 1) The development of a general framework based on half quadratic (HQ) minimization in combination with M-estimators in order to estimate the MDS embedding when the dissimilarity matrix is contaminated with outliers. 2) The proposal of three efficient algorithms, one based on the additive form of the HQ and another two resorting to the multiplicative form of the HQ, for finding the MDS solution. 3) The thorough study of the Welsch M-estimator, which is closely related to the maximum correntropy criterion, for solving the MDS problem when outliers are present. 4) The demonstration of the impact of various M-estimators in the solution of the MDS problem.

Throughout the paper the following notation is adopted: Scalars are denoted by lowercase letters (e.g., λ_1), vectors appear as lowercase boldface letters (e.g., $\mathbf{x}$), and matrices are denoted by uppercase boldface letters (e.g., $\mathbf{O}$). $(\cdot)^T$ denotes transposition, $\mathrm{tr}(\cdot)$ stands for the trace of the matrix inside parentheses, and the (i, j) element of $\mathbf{X}$ is represented by $[\mathbf{X}]_{ij}$ or x_{ij}. If $\mathbf{X}$ is a square matrix, then $\mathbf{X}^{-1}$ is its inverse. $\mathbf{I}$ stands for the identity matrix with compatible dimensions, $\mathrm{diag}(\mathbf{x})$ yields a square diagonal matrix with the elements of vector $\mathbf{x}$ appearing on its 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 $\mathbf{X}$ is declared by the row vector $\mathbf{x}^i$, while the j-th column is indicated with the column vector $\mathbf{x}_j$. The set of real and nonnegative real numbers is denoted by $\mathbb{R}$ and $\mathbb{R}_+$, respectively. Several norms of real-valued vectors and matrices are used. 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 double sum $\sum_{i=1}^{N}\sum_{j=i+1}^{N}(\cdot)$ is represented as $\sum_{i<j}^{N}(\cdot)$.

The remainder of this paper is structured as follows: The fundamentals of the MDS are surveyed in Section II. MDS techniques that reduce the influence of outliers are explored in Section III. Special emphasis is given to the Robust MDS (RMDS) proposed in [11]. An overview of M-estimators and their relation to correntropy is presented in Section IV. Section V deals with the additive and multiplicative forms of the HQ minimization. The proposed algorithms are detailed in Section VI. Section VII includes experimental results and a detailed comparison with the state-of-the-art techniques, demonstrating the merits of the proposed algorithms. Finally, Section VIII concludes the paper and provides pointers for further research.

II. MDS Overview

Let N denote the number of entities (objects) and d be their embedding dimension, e.g., 2 or 3. Let also $\mathbf{\Delta} = [\delta_{ij}]$ denote the pairwise dissimilarity matrix, where δ_{ij}, $i, j = 1, 2, ..., N$ refers to the dissimilarity between objects i and j. Such dissimilarities satisfy the nonnegativity and symmetry properties, i.e., a) $\delta_{ii} = \delta_{jj} = 0$, b) $\delta_{ij} \geq 0$, and c) $\delta_{ij} = \delta_{ji}$. The triangle inequality holds, if and only if the dissimilarities are distances. The derived embedding in a 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 the Hadamard product of $\mathbf{D}(\mathbf{X})$ with itself can be expressed as

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

where $\mathbf{E}$ is a $N \times N$ matrix of ones. The elements of $[\mathbf{D}(\mathbf{X})]^2$ are squared distances.

A least-squares (LS) loss function that measures the goodness of fit between δ_{ij} and $d_{ij}(\mathbf{X})$ is the raw stress defined as:

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

where w_{ij} is a nonnegative user-defined weight representing the importance of the dissimilarity between objects i and j. In most cases, all the weights are equal to one. Several schemes were proposed for employing unequal weights, such as the elastic scaling where $w_{ij} = \delta_{ij}^{-2}$ [12] and Sammon mapping where $w_{ij} = \delta_{ij}^{-1}$ [13]. The MDS is seeking $\mathbf{X}$ that minimizes (2). This is a non-convex optimization problem. Its solution is not unique, since rigid transformations (e.g., translation, rotation, or reflection) do not alter the value admitted by the stress function. Moreover, the axes arising by the application of MDS lack any physical interpretation.

A well known algorithm for solving MDS is the Scaling by Majorizing a Complicated Function (SMACOF) [14], where an iterative majorization of the stress function takes place. Prior to SMACOF, the popular classical MDS algorithm [1] and the gradient descent methods [3], [15] were applied. The majorization technique was later expanded in order to incorporate Minkowski distances [16], [17]. A survey of MDS can be found in [18], [19].

III. Robust Multidimensional Scaling

The fragility of any least-squares loss function (e.g., the stress) to outliers has motivated researchers to investigate alternatives that eliminate the influence of gross errors. Classical MDS and SMACOF techniques, despite their simplicity, are not robust, when the initial dissimilarities have been contaminated with outliers. Even a single outlier in the dissimilarity matrix $\mathbf{\Delta}$ may distort severely the solution of the classical MDS, because the noise is propagated to each element of the distance matrix through the double-centering process $-\frac{1}{2}\mathbf{J}\mathbf{\Delta}^2\mathbf{J}$, where $\mathbf{J} = \mathbf{I} - N^{-1}\,\mathbf{e}\,\mathbf{e}^T$ is the centering operator, $\mathbf{e}$ is a $N \times 1$ vector of ones, and $\mathbf{I}$ is the $N \times N$ identity matrix [20], [21]. Indeed, the classical MDS minimizes the loss function $\left\|-\frac{1}{2}\mathbf{J}[\mathbf{\Delta}^2 - \mathbf{D}^2]\mathbf{J}\right\|_F^2$. One could easily compensate for the effect of the double centering process by modifying the loss function of the classical MDS so as it does not incorporate the term $-\frac{1}{2}\mathbf{J}\mathbf{\Delta}^2\mathbf{J}$.

Alternatively, the cost function $\left\|\mathbf{\Delta}^2 - \mathbf{D}^2\right\|_1$, employed in the Robust Euclidean Embedding (REE) [21], could be used. A related idea was proposed in [22], i.e., $\sigma_1(\mathbf{X}) = \sum_{i<j}^{N} w_{ij}\,|\delta_{ij} - d_{ij}(\mathbf{X})|$. However, the ℓ_1 norm is not smooth, due to its singularity at its origin. To alleviate this problem, the Huber loss

function, which belongs to the broad class of M-estimators, was proposed [23]:

$$\sigma_H(\mathbf{X}) = \sum_{i<j}^{N} w_{ij}\, \phi_H(\delta_{ij} - d_{ij}(\mathbf{X})) \tag{3}$$

where

$$\phi_H(r) = \begin{cases} \frac{r^2}{2} & if \quad |r| \leq a \\ a|r| - \frac{a^2}{2} & if \quad |r| > a \end{cases} \tag{4}$$

and a is the threshold that can be chosen arbitrarily or adaptively from the data.

Let us assume that each dissimilarity 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 and o_{ij} models an outlier. Due to the sparseness of the outliers, a small amount of them is expected to admit a non-zero value. Accordingly, the inclusion of the ℓ_1 norm of the $N \times N$ outlier matrix $\mathbf{O}$ in the MDS loss function is fully justified, yielding [11]:

$$(\hat{\mathbf{O}}, \hat{\mathbf{X}}) = \underset{\mathbf{O},\mathbf{X}}{\operatorname{argmin}} \left\{ \sum_{i<j}^{N} (\delta_{ij} - d_{ij}(\mathbf{X}) - o_{ij})^2 + \lambda_1 \sum_{i<j}^{N} |o_{ij}| \right\}. \tag{5}$$

The first term in (5) corresponds to the goodness of fit between δ_{ij} and $d_{ij}(\mathbf{X})$ after subtracting the impact of outliers. The second term is a penalty related to sparsity requirement for $\mathbf{O}$, where λ_1 is a regularization parameter. By finding a majorizer function of the ℓ_1-norm regularized stress in (5) and implementing a Majorization-Minimization algorithm applied to the mazorizer with regard to $\mathbf{O}$ and $\mathbf{X}$ separately, the solution of (5) is given by the iterative procedure [11]:

$$o_{ij}^{(t+1)} = S_{\lambda_1}(\delta_{ij} - d_{ij}(\mathbf{X}^{(t)})) \tag{6}$$

$$\mathbf{X}^{(t+1)} = \mathbf{L}^{\dagger}\mathbf{L}_{+}(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)} \tag{7}$$

where $S_\lambda(x) = \operatorname{sign}(x)(|x| - \frac{\lambda}{2})_+$ 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 $[\mathbf{L}]_{ij} = -1$. Its rank is $N - 1$. Accordingly, $\mathbf{L}$ is not invertible, being not full rank. For this reason, the Moore-Penrose pseudoinverse is used in (7), which is defined as $\mathbf{L}^{\dagger} = N^{-1}\mathbf{J}$. In (7), the $\mathbf{L}_{+}(\mathbf{O}, \mathbf{X})$ is the Laplacian

matrix having elements:

$$[\mathbf{L}_+(\mathbf{O},\mathbf{X})]_{ij} = \begin{cases} -(\delta_{ij} - o_{ij})\, d_{ij}^{-1}(\mathbf{X}) & (i,j) \in \mathbb{S} \\ 0 & (i,j) \in \mathbb{T} \\ -\sum_{k=1,k\neq i}^{N} [\mathbf{L}_+(\mathbf{O},\mathbf{X})]_{ik} & (i,j) \in \mathbb{Q} \end{cases} \tag{8}$$

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 just described algorithm was coined as Robust MDS (RMDS). The initial configuration $\mathbf{X}^{(0)}$ is chosen randomly, while the initial outlier matrix $\mathbf{O}^{(0)}$ is set to zero.

Given $\mathbf{X}^{(t)}$, the estimation of $\mathbf{O}^{(t+1)}$ via (6) constitutes a ℓ_1 regularization (LASSO) problem. Given $\mathbf{O}^{(t)}$, (7) is essentially the least-squares solution of the system of the equations $\mathbf{L}\mathbf{X}^{(t+1)} = \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})$ $\mathbf{X}^{(t)}$. That is, it is the solution of the optimization problem $\left\|\mathbf{L}\mathbf{X}^{(t+1)} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}\right\|_F^2$. It has been attested by extensive experimentation that (5) is still vulnerable to outliers, even though it alleviates their impact. In highly contaminated environments, the RMDS cannot yield an acceptable approximation of the initial configuration for any λ_1 value in most of the cases.

Taking into account the aforementioned ascertainment that (7) is a LS solution and the certitude that a LS problem is strongly influenced by outliers, it is proposed to substitute the squared Frobenius norm yielding (7) with an M-estimator which downweights the impact of gross errors due to outliers. More precisely, it is proposed to seek 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 smoothness regularization term through the Frobenius norm of $\mathbf{X}$, i.e.,

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ \phi(\mathbf{L}\mathbf{X} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}) + \lambda_2 \left\|\mathbf{X}\right\|_F^2 \right\}. \tag{9}$$

The adoption of M-estimators pursues to mitigate the outliers effect further since even if $\mathbf{O}^{(t+1)}$ estimation is not robust, $\mathbf{X}^{(t+1)}$ will be not so sensitive to inaccurate $\mathbf{O}^{(t+1)}$ calculation. Essentially, the solution of the optimization problem (9) is proposed as an principled alternative to substitute (7) within the RMDS algorithm [11].

To summarize, this work addresses the major limitations of RMDS due to the LS viewpoint yielding

(7) and the lack of any smoothness term. Moreover, the optimization problem (9) proposed here is solved in a principled way by HQ minimization. Links are established with the maximum correntropy criterion, which is strongly related to the Welsch M-estimator, to increase cohesion. By doing so, a unified framework emerges that extends the work in [11].

IV. M-ESTIMATORS AND CORRENTROPY

Let r_i denote residuals that depend on parameter x. M-estimators minimize a loss function $\sum_{i=1}^{N} \phi(r_i; x)$ with respect to x, i.e., $\hat{x} = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{N} \phi(r_i; x)$, where $\phi(r; x)$ is called *potential function* [24]. M-estimators aim at the minimization of bias due to outliers by replacing the least-squares loss function, that constitutes a special case of M-estimators being susceptible to outliers, with another function that increases less than the squared error and thus being less fragile to gross errors. Potential functions for a variety of M-estimators are detailed in the next Section. Beyond convex M-estimators (like the ℓ_2, ℓ_1, ℓ_p, ℓ_1-ℓ_2, log-cosh, Huber, and Fair estimators), there are also non-convex M-estimators, such as the Cauchy, Geman-McClure, Welsch, and Tukey estimators. Frequently, the non-convex M-estimators are more efficient than the corresponding convex ones [25]. The properties of potential functions can be found in [26].

The (cross) correntropy, first introduced as a generalized correlation function [27], 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 σ denoting its size [26]. It is well known that the joint probability density function (pdf) can not be accurately estimated when a finite amount of data (y_i, w_i), $i = 1, 2, ..., N$ is available. In such a case, the sample estimator of correntropy is recommended [26], i.e., $\hat{V}_\sigma(W, Y) = \frac{1}{N} \sum_{i=1}^{N} g_\sigma(w_i - y_i)$.

The correntropy measure is symmetric, positive, and bounded. The maximum is achieved at $W = Y$. All correntropy properties depend on the kernel size, which is application specific [26]. 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 [26]

$$CIM(\underline{W}, \underline{Y}) = \left[g_\sigma(0) - \frac{1}{N} \sum_{i=1}^{N} g_\sigma(w_i - y_i) \right]^{1/2}. \tag{10}$$

The CIM possesses the properties of symmetry, non-negativity and triangle inequality [26]. Additionally,

the identity of indiscernibles holds, i.e., $CIM(\underline{W},\underline{Y}) = 0$, if and only if $\underline{W} = \underline{Y}$ [26]. The Maximum Correntropy Criterion (MCC) aims at maximizing the sample correntropy (i.e., the last term in (10)). Since CIM is a decreasing function of correntropy, the maximization of correntropy is equivalent to the minimization of the CIM.

It is seen that $E[g_\sigma(W-Y)]$ resembles the mean squared error $MSE = E[(W-Y)^2]$ for $g_\sigma(W-Y) = (W-Y)^2$. The Gaussian kernel function makes the MCC a local criterion, while the MSE is a global one [26]. The term global implies that all sample errors conduce significantly to the estimation of MSE. On the contrary, the Gaussian kernels restrict the analysis to a local region of the joint space. Indeed, the correntropy depends heavily on the kernel function along the line $w = y$. The correntropy is closely related to M-estimators [26]. By setting $\phi(x) = 1 - g_\sigma(x)$, the CIM becomes equivalent to the Welsch M-estimator [28]. The MCC, as a similarity metric, has proven to be appropriate in non-linear, non Gaussian signal processing applications, such as robust regression [26], feature selection [29], etc.

V. Half-Quadratic Minimization

Next, let us briefly describe the half-quadratic minimization for a scalar function. A new objective function is introduced that depends both on the initial variable x and a new auxiliary variable p. Precisely, if $\mathcal{J}(x)$ is the initial objective function and $J(x,p)$ is the new objective function, then one requires $\mathcal{J}(x) = \min_p\{J(x,p)\}$, $\forall x$. If p is fixed, J is quadratic w.r.t. x. Hence, the name *Half Quadratic*. The global minimum of the new objective function w.r.t. x is the same with that of the initial objective function. However, the estimation of the argument x becomes considerably easier, because of the specific formulation by which the auxiliary variable p is initiated and the alternating minimization procedure that takes place. Essentially, the HQ theory is based on the estimation of alternating updates of p and x. For simplicity, let $J(x,p) = h(x) + \beta\, Q(\gamma\, x, p) + \psi(p)$, where $h(x)$ is quadratic w.r.t. x, $Q(\cdot,p)$: $\mathbb{R} \to \mathbb{R}$ is quadratic for any $p \in \mathbb{R}$, and $\psi(\cdot)$: $\mathbb{R} \to \mathbb{R}$ satisfies [25]:

$$\phi(x) = \min_p\{Q(x,p) + \psi(p)\} \quad \forall x \in \mathbb{R} \tag{11}$$

where $\phi(\cdot)$ is a potential function either convex or non-convex. (11) implies that $\psi(\cdot)$ is the *conjugate function* of $\phi(\cdot)$ (see further [30, ch. 3, p. 90]). The auxiliary variable p is determined by the HQ minimizer

function $\delta(\cdot)$ derived by $\psi(\cdot)$ and thus related to $\phi(\cdot)$. The minimizer function satisfies the constraint $Q(x,\delta(x))+\psi(\delta(x)) \leq Q(x,p)+\psi(p)$, $\forall p \in \mathbb{R}$ [31].

The loss function $\phi(\cdot)$ may be one of M-estimators, while $Q(x,p)$ is a quadratic function admitting two forms. Namely, the multiplicative form

$$Q_M(x,p) = px^2 \qquad p \in \mathbb{R}_+, \quad x \in \mathbb{R} \tag{12}$$

which results to the loss function $\phi(x) = \min_p \{px^2 + \psi(p)\}$ [32] and the additive form, proposed in [33]:

$$Q_A(x,p) = (x\sqrt{c} - \frac{p}{\sqrt{c}})^2 \qquad p \in \mathbb{R}, \quad x \in \mathbb{R} \tag{13}$$

which results to the loss function $\phi(x) = \min_p \{(x\sqrt{c} - \frac{p}{\sqrt{c}})^2 + \psi(p)\}$, where c is a positive constant. In both forms of the HQ minimization $\phi(x)$ should fulfil certain conditions [31].

The minimizer function $\delta(\cdot)$ is called *weighting function*. $\delta(x)$ admits distinct additive and multiplicative formulations. For $\phi(x)$: $\mathbb{R} \rightarrow \mathbb{R}$, these formulations are [31]:

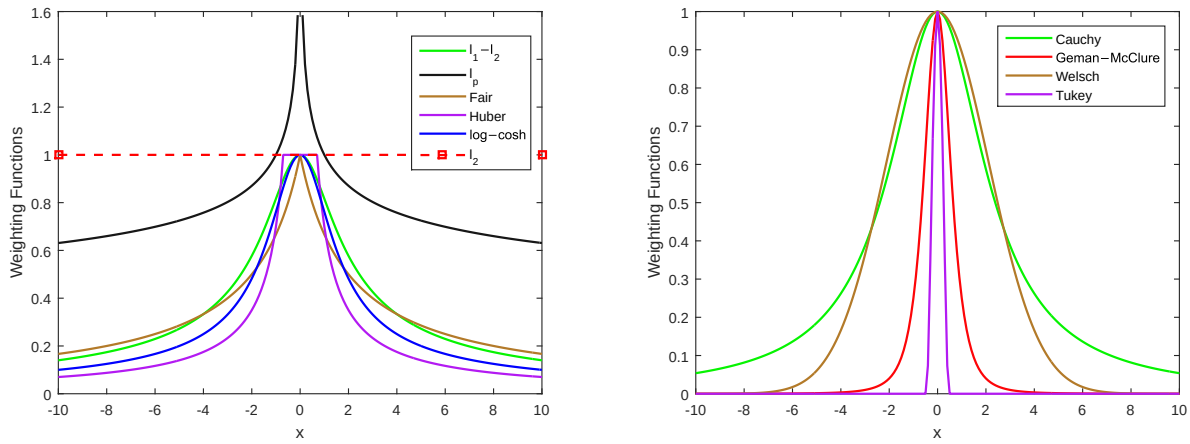
$$\delta_A(x) = cx - \phi'(x) \tag{14}$$

$$\delta_M(x) = \begin{cases} \phi''(0^+) & \text{if } x = 0 \\ \frac{\phi'(x)}{x} & \text{if } x \neq 0. \end{cases} \tag{15}$$

It is proven [31] that the optimal value of the positive constant c is $c = \sup_{x\in\mathbb{R}} \phi''(x)$. Taking into account that for most M-estimators $\phi''(0) = \sup_{x\in\mathbb{R}} \phi''(x)$, it follows that $c = \phi''(0)$.

As mentioned previously, the potential (or loss) function $\phi(x)$: $\mathbb{R} \rightarrow \mathbb{R}$ can be associated to an M-estimator. Table I summarizes the various potential functions $\phi(x)$ and their corresponding weighting functions $\delta(x)$: $\mathbb{R} \rightarrow \mathbb{R}$ for the additive and multiplicative form of the HQ. Note that not all potential functions $\phi(x)$ fulfil the conditions in [31]. Various weighting functions $\delta(x)$ for convex and non-convex potential functions in the multiplicative form of HQ are plotted in Figures 1a and 1b.

The augmented function $J(x,p)$: $\mathbb{R} \rightarrow \mathbb{R}$ is non-increasing at each iteration. Let us denote the solutions found at iteration t as $(x^{(t)}, p^{(t)})$. A basic property of the HQ minimization is $J(x^{(t+1)}, p^{(t+1)}) \leq J(x^{(t)}, p^{(t+1)}) \leq J(x^{(t)}, p^{(t)})$ [31]. Furthermore, the convergence of the sequence $(\ldots, J(x^{(t)}, p^{(t)}), J(x^{(t)}, p^{(t+1)}), J(x^{(t+1)}, p^{(t+1)}), \ldots)$ is guaranteed, when HQ is employed [31]. Although the multiplicative form



(a) Convex potential functions

(b) Non-convex potential functions

Fig. 1: Weighting functions for convex and non-convex potential functions for the multiplicative form of HQ.

requires fewer iterations than the additive form to converge, the computational cost at each iteration is heavier [31].

TABLE I: Potential 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	$\phi(x) = \begin{cases} \frac{a^2}{6}(1-[1-(\frac{x}{a})^2]^3) & \lvert x\rvert \le a \\ \frac{a^2}{6} & \lvert x\rvert > a \end{cases}$	$\delta(x) = \begin{cases} [1-(\frac{x}{a})^2]^2 & \lvert x\rvert \le a \\ 0 & \lvert x\rvert > a \end{cases}$	$\delta(x) = \begin{cases} cx - x[1-(\frac{x}{a})^2]^2 & \lvert x\rvert \le a \\ 0 & \lvert x\rvert > a \end{cases}$

VI. A Unified Half-Quadratic Framework for MDS When Outliers are Present

In this section, the optimization problem (9) is solved with HQ minimization, extending the analysis of Section V to multivariate functions. Let $\phi(\mathbf{x}) = \sum_{i=1}^{N} \phi(x_i)$ be a loss function on a vector $\mathbf{x} \in \mathbb{R}^{N\times 1}$, where x_i is the i-th entry of $\mathbf{x}$. The minimization problem (9) can take the form:

$$(\hat{\mathbf{x}}, \hat{\mathbf{p}}) = \underset{\mathbf{x},\mathbf{p}}{\operatorname{argmin}}\{J(\mathbf{x},\mathbf{p})\} = \underset{\mathbf{x},\mathbf{p}}{\operatorname{argmin}}\Big\{Q(\mathbf{x},\mathbf{p}) + \sum_{i=1}^{N} \psi(p_i) + h(\mathbf{x})\Big\}. \tag{16}$$

The solution $(\hat{\mathbf{x}}, \hat{\mathbf{p}})$ of the optimization problem (16) can be obtained in an alternating fashion as follows:

$$\mathbf{p}^{(t+1)} = \delta(\mathbf{x}^{(t)}) \tag{17}$$

$$\mathbf{x}^{(t+1)} = \underset{\mathbf{x}}{\operatorname{argmin}}\{Q(\mathbf{x},\mathbf{p}^{(t+1)}) + h(\mathbf{x})\}. \tag{18}$$

For $\phi(\mathbf{x})$: $\mathbb{R}^{N\times 1} \to \mathbb{R}$, (14) and (15) are applied componentwise. The convex penalty function $h(\mathbf{x})$ can be defined as $h(\mathbf{x}) = \lambda_2 \|\mathbf{x}\|_1$, $h(\mathbf{x}) = \lambda_2 \|\mathbf{x}\|_2^2$, or $h(\mathbf{x}) = \lambda_2 \|\mathbf{x}\|_{2,1}$. In the following, we derive the solutions of (17) and (18) in the context of (9).

A. Additive Form

The quadratic function $Q_A(\cdot)$ of the additive form of the HQ is defined as

$$Q_A(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}, \mathbf{P}) = \left\| \sqrt{c}\,(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) - \frac{1}{\sqrt{c}}\,\mathbf{P} \right\|_F^2 \tag{19}$$

where $\mathbf{P} \in \mathbb{R}^{N\times d}$ is a matrix of auxiliary variables, determined by the minimizer function $\delta_A(\cdot)$. The potential loss function $\phi_A(\cdot)$ is defined as:

$$\phi_A(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) = \min_{\mathbf{P}}\Big\{ Q_A(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}, \mathbf{P}) + \sum_{n=1}^{N}\sum_{m=1}^{d} \psi(p_{nm}) \Big\} \tag{20}$$

where $\psi(\cdot)$ is the conjugate function of $\phi_A(\cdot)$. Accordingly, $J_A(\mathbf{X},\mathbf{P})$ in (16) is given by:

$$J_A(\mathbf{X},\mathbf{P}) = \left\| \sqrt{c}\,(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) - \frac{1}{\sqrt{c}}\,\mathbf{P} \right\|_F^2 + \sum_{n=1}^{N}\sum_{m=1}^{d} \psi(p_{nm}) + \lambda_2 \|\mathbf{X}\|_F^2 \tag{21}$$

where λ_2 is a positive parameter that regulates the Frobenius norm of $\mathbf{X}$. The estimation of $(\mathbf{X},\mathbf{P})$ can be derived by solving $(\hat{\mathbf{X}},\hat{\mathbf{P}}) = \underset{\mathbf{X},\mathbf{P}}{\operatorname{argmin}}\{J_A(\mathbf{X},\mathbf{P})\}$. When the solution for $\mathbf{X}$ is sought, the terms including

$\psi(\cdot)$ can be omitted due to the fact that the auxiliary variables depend only on the minimizer function $\delta_A(\cdot)$, as indicated in (17), 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:

$$\mathbf{P}^{(t+1)} = \delta_A(\mathbf{L}\mathbf{X}^{(t)} - \mathbf{Y}) \tag{22}$$

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ \left\| \sqrt{c}\,(\mathbf{L}\mathbf{X} - \mathbf{L}_+\mathbf{X}^{(t)}) - \frac{1}{\sqrt{c}}\,\mathbf{P}^{(t+1)} \right\|_F^2 + \lambda_2 \left\|\mathbf{X}\right\|_F^2 \right\}. \tag{23}$$

The optimization problem (23) can be reformulated as:

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ c\,\operatorname{tr}((\mathbf{L}\mathbf{X} - \mathbf{H}^{(t+1)})^T(\mathbf{L}\mathbf{X} - \mathbf{H}^{(t+1)})) + \lambda_2\,\operatorname{tr}(\mathbf{X}^T\mathbf{X}) \right\}. \tag{24}$$

where

$$\mathbf{H}^{(t+1)} = \mathbf{Y} + \frac{1}{c}\,\mathbf{P}^{(t+1)}. \tag{25}$$

By applying the first order optimality condition to (24) w.r.t. $\mathbf{X}$, a closed form solution for $\mathbf{X}^{(t+1)}$ is obtained[1]:

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

The objective function is minimized at each iteration until its convergence. In this form of the HQ, the auxiliary variables $\mathbf{P}$ can be viewed as errors incurred by noise. At each iteration, outlying observations are adjusted gradually, because the loss function $\phi_A(\cdot)$ corresponds to an M-estimator. The complete procedure for the solution of (9) by the additive form of the HQ minimization is outlined in the Algorithm 1. The initial configuration $\mathbf{X}^{(0)}$ can be chosen randomly or can be set to the solution of the classical MDS algorithm. The initial outlier matrix $\mathbf{O}^{(0)}$ is set to zero.

B. Multiplicative Form (Version 1)

For the multiplicative form of the HQ, the quadratic function $Q_M(\cdot)$ is defined as

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

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

Algorithm 1 Additive Form of the HQ Minimization for MDS (HQAMDS)

Input: Initial outlier matrix $\mathbf{O}^{(0)}$ and initial configuration $\mathbf{X}^{(0)}$
Output: Outlier matrix $\mathbf{O}^{(t+1)}$ and coordinate matrix $\mathbf{X}^{(t+1)}$

```
for t = 0, 1, 2, ... do
    Find each entry of O^(t+1) via (6)
    Update P^(t+1) via (22) with L_+ as in (8)
    Update H^(t+1) via (25) with L_+ as in (8)
    Update X^(t+1) via (26)
end for
```

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 (15). It is seen that $Q_M(\cdot)$ is the weighted sum of squared ℓ_2 norms of the rows of the residual matrix $\mathbf{LX} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$. The potential loss function $\phi_M(\cdot)$ is defined as

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

Using (28), the objective function in (16) takes the form

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

Let $(\hat{\mathbf{X}}, \hat{\mathbf{p}}) = \underset{\mathbf{X},\mathbf{p}}{\operatorname{argmin}}\{J_M(\mathbf{X}, \mathbf{p})\}$. Due to the fact that the auxiliary variables in (17) depend only on the minimizer function $\delta_M(\cdot)$, the terms $\psi(\cdot)$ can be omitted, because the auxiliary variables are fixed, when we minimize w.r.t to $\mathbf{X}$. Thus, a local minimizer $(\hat{\mathbf{X}}, \hat{\mathbf{p}})$ can be estimated using the alternating minimization:

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

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ \operatorname{tr}((\mathbf{LX} - \mathbf{Y})^T \mathbf{P}^{(t+1)} (\mathbf{LX} - \mathbf{Y})) + \lambda_2 \operatorname{tr}(\mathbf{X}^T\mathbf{X}) \right\} \tag{31}$$

where $\mathbf{P}^{(t+1)} = \operatorname{diag}(\mathbf{p}^{(t+1)})$ is a diagonal matrix with ii-th element equal to $p_i^{(t+1)}$. A pertinent approach was proposed in [29], focusing on robust feature selection. The optimization problem (31)

can be reformulated as a least-squares regression problem:

$$\mathbf{X}^{(t+1)} = \underset{\mathbf{X}}{\operatorname{argmin}} \left\{ \left\| \tilde{\mathbf{L}}\mathbf{X} - \tilde{\mathbf{Y}} \right\|_F^2 + \lambda_2 \left\| \mathbf{X} \right\|_F^2 \right\} \tag{32}$$

where $\tilde{\mathbf{L}} = \sqrt{\mathbf{P}^{(t+1)}}\mathbf{L}$ and $\tilde{\mathbf{Y}} = \sqrt{\mathbf{P}^{(t+1)}}\mathbf{Y}$. Setting the derivative of (31) w.r.t to $\mathbf{X}$ equal to zero, the following closed-form solution is obtained:

$$\mathbf{X}^{(t+1)} = (\mathbf{L}^T\mathbf{P}^{(t+1)}\mathbf{L} + \lambda_2\mathbf{I})^{-1}\mathbf{L}^T\mathbf{P}^{(t+1)}\mathbf{Y}. \tag{33}$$

At each iteration, the auxiliary variable p_i represents the weight that regulates the impact of $\|(\mathbf{LX} - \mathbf{L}_+ \mathbf{X}^{(t)})^i\|_2$. The introduction of M-estimators into the augmented objective function reduces the influence of the outliers, since $p_i^{(t+1)}$ always admits a low weight, as manifested by the presence of $\delta_M(\cdot)$ in (30) that is associated to the potential function $\phi_M(\cdot)$ of an M-estimator. The multiplicative form 1 of the HQ minimization is essentially an iterative reweighed least-squares (IRLS) minimization. The complete procedure for the solution of (9) by using this version of the multiplicative form of HQ is outlined in Algorithm 2. Initialization can be done as in Subsection A.

C. Multiplicative Form (Version 2)

Alternatively, the quadratic function $Q_M(\cdot)$ can be defined as the weighted sum of all squared elements of $\mathbf{LX} - \mathbf{L}_+(\mathbf{O}^{(t+1)}, \mathbf{X}^{(t)})\mathbf{X}^{(t)}$:

$$Q_M(\mathbf{LX} - \mathbf{L}_+\mathbf{X}^{(t)}, \mathbf{P}) = \sum_{i=1}^{N}\sum_{j=1}^{d} p_{ij} \left[\mathbf{LX} - \mathbf{L}_+\mathbf{X}^{(t)} \right]_{ij}^2 \tag{34}$$

where $\mathbf{P} \in \mathbb{R}^{N\times d}$. The potential loss function $\phi_M(\cdot)$ associated to (34) is given by

$$\phi_M(\mathbf{LX} - \mathbf{L}_+\mathbf{X}^{(t)}) = \min_{\mathbf{P}} \left\{ \sum_{i=1}^{N}\sum_{j=1}^{d} p_{ij} \left[\mathbf{LX} - \mathbf{L}_+\mathbf{X}^{(t)} \right]_{ij}^2 + \sum_{i=1}^{N}\sum_{j=1}^{d} \psi(p_{ij}) \right\}. \tag{35}$$

Using (35), the objective function in (16) is rewritten as

$$J_M(\mathbf{X}, \mathbf{P}) = \phi_M(\mathbf{LX} - \mathbf{L}_+\mathbf{X}^{(t)}) + \lambda_2 \left\| \mathbf{X} \right\|_F^2. \tag{36}$$

The main difference between (29) and (36) is the incorporation of all individual residuals $[\mathbf{LX}-\mathbf{Y}]_{ij}$ instead of the ℓ_2 norm for each row $(\mathbf{LX}-\mathbf{Y})^i$. A related approach was proposed in [34], focusing on robust subspace clustering problem. The function $J_M(\mathbf{X},\mathbf{P})$ can be minimized in an alternating fashion as follows:

$$p_{ij}^{(t+1)} = \delta_M\left([\mathbf{LX}^{(t)}-\mathbf{Y}]_{ij}\right) \tag{37}$$

$$\mathbf{x}_j^{(t+1)} = \underset{\mathbf{x}_j}{\operatorname{argmin}}\left\{(\mathbf{Lx}_j-\mathbf{y}_j)^T\mathbf{P}_j^{(t+1)}(\mathbf{Lx}_j-\mathbf{y}_j)+\lambda_2\left\|\mathbf{x}_j\right\|_2^2\right\} \tag{38}$$

where $\mathbf{y}_j = \mathbf{L}_+\mathbf{x}_j^{(t)}$ and $\mathbf{P}_j^{(t+1)}$ is a diagonal matrix with (i,i)-th element $[\mathbf{P}_j^{(t+1)}]_{ii} = p_{ij}^{(t+1)}$. The optimization problem (38) can be transformed into a least-squares regression problem

$$\mathbf{x}_j^{(t+1)} = \underset{\mathbf{x}_j}{\operatorname{argmin}}\left\{\left\|\tilde{\mathbf{L}}\mathbf{x}_j-\tilde{\mathbf{y}}_j\right\|_2^2+\lambda_2\left\|\mathbf{x}_j\right\|_2^2\right\} \tag{39}$$

where $\tilde{\mathbf{L}} = \sqrt{\mathbf{P}_j^{(t+1)}}\mathbf{L}$ and $\tilde{\mathbf{y}}_j = \sqrt{\mathbf{P}_j^{(t+1)}}\mathbf{y}_j$. By applying the first order optimality condition to (38) w.r.t to $\mathbf{x}_j$, the following closed form results:

$$\mathbf{x}_j^{(t+1)} = (\mathbf{L}^T\mathbf{P}_j^{(t+1)}\mathbf{L}+\lambda_2\mathbf{I})^{-1}\mathbf{L}^T\mathbf{P}_j^{(t+1)}\mathbf{y}_j. \tag{40}$$

The complete procedure for the solution of (9) by using this version is outlined in Algorithm 2. The initial matrices can be set as said previously.

Algorithm 2 Multiplicative Forms of the HQ Minimization for the MDS (HQMMDS1 and HQMMDS2)

Input: Initial outlier matrix $\mathbf{O}^{(0)}$ and initial configuration $\mathbf{X}^{(0)}$
Output: Outlier matrix $\mathbf{O}^{(t+1)}$ and coordinate matrix $\mathbf{X}^{(t+1)}$

```
for t = 0, 1, 2, ... do
    Find each entry of O^(t+1) via (6)
    if version 1 then
        Update p_i^(t+1) via (30) with L_+ as in (8)
        Update X^(t+1) via (33)
    else if version 2 then
        Update p_ij^(t+1) via (37) with L_+ as in (8)
        Update x_j^(t+1) via (40)
    end if
end for
```

VII. Numerical Tests

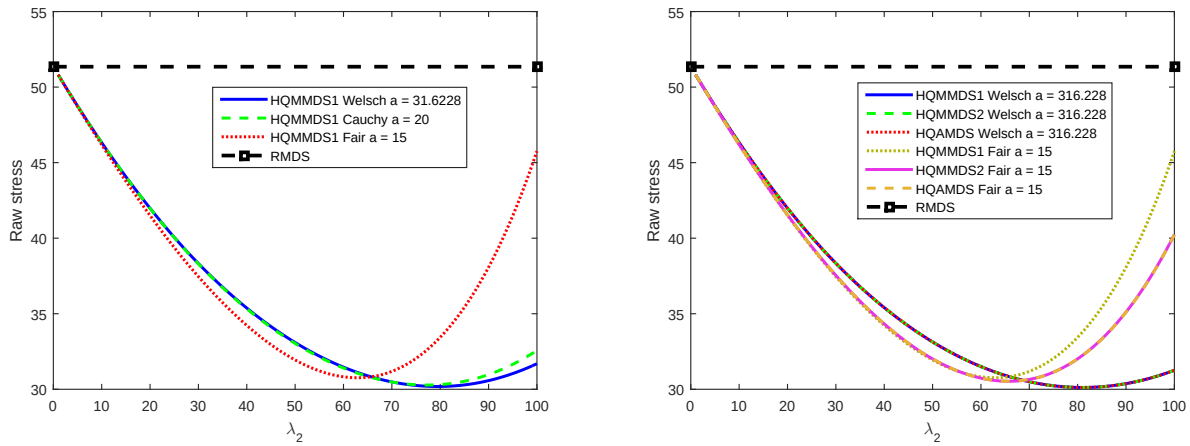
The additive and the multiplicative forms of the HQ minimization for the MDS were implemented in Matlab and tested on several dissimilarity matrices $\mathbf{\Delta}$. In order to evaluate and benchmark the proposed algorithms, three additional MDS techniques were implemented in the same environment and tested on the same dissimilarity matrices. These techniques were: a) the popular SMACOF algorithm [14], reported to be one of the most efficient algorithms, b) the subgradient version of REE algorithm [21], and c) the RMDS [11]. For all techniques, authors' recommendations were strictly followed, while the implementation was intended to achieve the best possible performance.

The quality of the embedding for each algorithm was evaluated with respect to four figures of merit: a) the normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}}) = \sqrt{\frac{\sum_{(i,j)\in\mathbb{U}}(\delta_{ij} - d_{ij}(\mathbf{X}))^2}{\sum_{(i,j)\in\mathbb{U}}\delta_{ij}^2}}$, where $\mathbb{U}$ denotes the set of outlier-free dissimilarities, namely when $[\mathbf{O}]_{ij} = 0$, as in [11]; b) the number of estimated outliers $\hat{S}$, as in [11]; c) the distortion (raw stress) $\sigma_r(\hat{\mathbf{X}})$, defined in (2), between the derived embedding and the noise-free configuration; and d) the standardized Procrustean goodness-of-fit criterion ϱ defined as the sum of the squared errors standardized by a measure of the scale $\mathbf{X}$[2]. The last criterion can only be applied to fixed configurations and assesses the linear transformation of an embedding against the points of the fixed configuration. The MDS is highly related with Procrustean techniques, whose objective is to transform an initial configuration, which can be an MDS solution, to a target configuration, as closely as possible [19]. These techniques unveil the coherence between the different MDS solutions and serve as an important tool for rejecting deceptive and inappropriate MDS embeddings.

To assess the aforementioned methods, 100 Monte Carlo simulations of RMDS algorithm were run, using a different random initial configuration $\mathbf{X}^{(0)}$ in each run. From all runs, the reported figures of merit refer to the case where RMDS algorithm has exhibited the minimum value in the raw stress $\sigma_r(\hat{\mathbf{X}})$, namely when the derived embedding was closer to the noise-free configuration. RMDS, HQAMDS, and HQMMDS algorithms terminated when the fraction $\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_F / \left\|\mathbf{X}^{(t+1)}\right\|_F$ was less than 10^{-6} or when the number of iterations reached 5000.

An important aspect of MDS algorithms is their initialization. In the majority of the experiments, the

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



(a) HQMMDS1 raw stress $\sigma_r(\hat{\mathbf{X}})$ (b) HQAMDS and HQMMDS raw stress $\sigma_r(\hat{\mathbf{X}})$

Fig. 2: HQAMDS, HQMMDS1 and HQMMDS2 embedding quality on the first data set for $\varpi = 12\%$.

classical MDS solution [1] was proven to be a worse initialization than a random configuration. All algorithms were tested on four data sets. The first data set was chosen to be a fixed configuration. Three real data sets were also employed.

A. First Data Set

The first data set comprises a square with $N = 100$ points in the two-dimensional space. The bottom-left point is at $(1, 1)$, the upper-right point is at $(10, 10)$, while all points are equidistant from their vertical and horizontal neighbors by one unit. Each element of the initial dissimilarity matrix was contaminated with a background error ϵ_{ij} derived from a zero mean truncated Gaussian distribution with variance 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 $\varpi\%$ was set at $594/(100 \cdot 99/2) = 12\%$. Another way of selecting the outlier indices is to shuffle randomly the elements in $\mathbf{\Delta}$ and then select the first $\varpi\%$ elements that will be contaminated with outliers [20].

Let a_h be the parameter of the Huber M-estimator and $\hat{\sigma}_\epsilon$ be the median absolute deviation (MAD)[3] of nominal errors. Taking into account the equivalence with Huber M-estimator for $\lambda_1 = 2a_h$ and that

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

$a_h = 1.345 \times 1.483 \times \hat{\sigma}_\epsilon$ yields 95% asymptotic efficiency for the normal distribution [35], λ_1 was set to $3.98927\,\hat{\sigma}_\epsilon$ for both RMDS and HQMMDS.

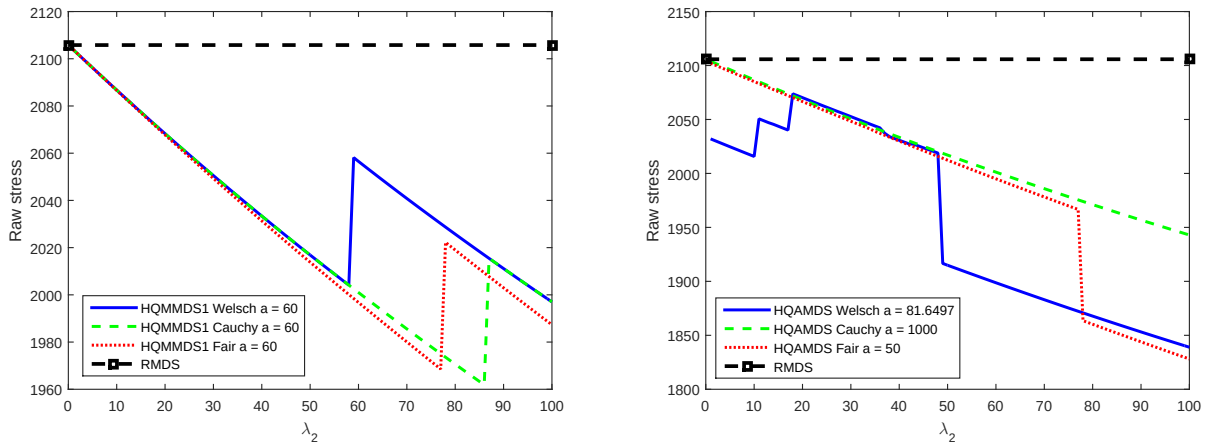
Table II gathers the figures of merit related to the embedding quality delivered by SMACOF, REE, and RMDS. The reported figures of REE algorithm were measured after 4000 iterations. Due to lack of space, only the raw stress $\sigma_r(\hat{\mathbf{X}})$ of the three proposed algorithms is plotted in Figures 2a-2b for $\lambda_2 \in [1, 100]$. More specifically, $\sigma_r(\hat{\mathbf{X}})$ for HQMMDS1 is shown in Figure 2a with a being set to 31.6228, 20, and 15 for the Welsch, Cauchy, and Fair M-estimators. $\sigma_r(\hat{\mathbf{X}})$ is plotted for HQAMDS, HQMMDS1 and HQMMDS2 in Figure 2b for Welsch M-estimator for $a = 316.228$. The plots of the same figure of merit for the Fair M-estimator in HQAMDS, HQMMDS1 and HQMMDS2 are overlaid in Figure 2b for $a = 15$. In HQAMDS, parameter c was equal to 1. The plots of $\hat{S}$ and ϱ for the aforementioned M-estimators in Figures 2a-2b look similar to $\sigma_r(\hat{\mathbf{X}})$ and are always smaller than RMDS for $\lambda_2 \in [1, 100]$. Generally, $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ for these M-estimators admits smaller values than RMDS for $\lambda_2 \in [1, 50]$, which indicates that $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ without $\hat{S}$ or $\sigma_r(\hat{\mathbf{X}})$ is not a reliable figure of merit for judging the quality of an embedding.

It can be seen in Figure 2b that the additive form and both versions of the multiplicative form for the same parameter a yield similar results provided that a admits a large value (e.g., $a = 316.228$) so that the Welsch M-estimator approximates the ℓ_2 estimator. On the contrary, when a admits a small value (e.g., $a = 15$ for Fair M-estimator), it appears that $\sigma_r(\hat{\mathbf{X}})$ is affected by λ_2 values less in HQMMDS2 than HQMMDS1.

TABLE II: Figures of merit for the embedding quality obtained by SMACOF, REE, and RMDS applied to the 1st data set.

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

It is self-evident that when the aforementioned M-estimators are employed in HQAMDS, HQMMDS1 and HQMMDS2, the resulting embedding outperforms the one derived by RMDS with respect to $\sigma_r(\hat{\mathbf{X}})$



(a) HQMMDS1 raw stress $\sigma_r(\hat{\mathbf{X}})$ (b) HQAMDS raw stress $\sigma_r(\hat{\mathbf{X}})$

Fig. 3: HQMMDS1 and HQAMDS embedding quality on the 2nd data set for $\varpi = 15.82\%$.

for a wide range of values admitted by λ_2. Needless to say that the same remark holds for the embeddings derived by SMACOF and REE.

B. Second Data Set

The second data set is composed by sequential employment records for $N = 80$ randomly selected Lloyds Bank employees from cohort 1925-1929 [36]. The data file contains 73 variables: an ID variable, a variable corresponding to the first year of employment (which is between 1925-1929), and 71 variables with the sequential data concerning career characteristics, as branch size, branch type, and job category [37]. An optimal matching algorithm, which estimates the minimum total cost of transforming one sequence into another between all potential transformations, is used to generate the dissimilarity matrix [37]. That is, the assessment of the difficulty for transforming the sequence i into sequence j is quantified by the dissimilarity δ_{ij}. The transformation of one career into another can encompass substitution, insertion, and deletion operations. The cost of each insertion or deletion is fixed, while substitution cost depends solely on the transformation pairs. In addition, the distances are standardized by the length of the longest career sequence. The data set was artificially contaminated by $500/(80 \cdot 79/2) = 15.82\%$ outliers, which were drawn from a uniform distribution on the range $[0, 3\max(\delta_{ij})]$. The outliers indices were chosen randomly.

λ_1 was set to 3.58 in order RMDS identifies $\hat{S} = 500$ outliers. The same value for λ_1 was used in both HQMMDS and HQAMDS. For HQMMDS1, the parameter a was set to 60 for the Welsch, Cauchy and Fair M-estimators. In HQAMDS, parameter c was equal to 1, while a was set to 81.6497, 1000, and 50 for the Welsch, Cauchy, and Fair M-estimators, respectively. It is worth noting that the parameter a of the Welsch M-estimator in HQAMDS was set to such a value so that HQAMDS with the Welsch M-estimator would converge to a local minimum of the raw stress $\sigma_r(\hat{\mathbf{X}})$ for a smaller λ_2 value than when HQAMDS with the Cauchy M-estimator was used. In the latter case, the parameter a of the Cauchy M-estimator was set to a large value so that its performance within HQAMDS is identical with that achieved by the ℓ_2 M-estimator.

TABLE III: Figures of merit for the embedding quality obtained by SMACOF, REE, and RMDS applied to the 2nd data set.

Outlier percentage $\varpi = 15.82\%$	SMACOF	REE	RMDS
Normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$	0.6045	0.6821	0.1971
Estimated outliers $\hat{S}$	-	-	500
Raw stress $\sigma_r(\hat{\mathbf{X}})$	14175.5	3017.7	2105.8

The figures of merit used to judge the embedding quality obtained by SMACOF, REE, and RMDS are summarized in Table III. The reported figures for REE were measured after 2000 iterations. The raw stress $\sigma_r(\hat{\mathbf{X}})$ of HQAMDS and HQMMDS1 is plotted in Figure 3 for $\lambda_2 \in [1, 100]$ and various M-estimators. For $\lambda_2 \in [1, 100]$, the plot of the normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$, in both forms, is roughly the same with that of $\sigma_r(\hat{\mathbf{X}})$ and is always smaller than that of RMDS. The estimated number of outliers $\hat{S}$, in both forms, was proven to be relatively constant, near the value of 500. It is apparent that the proposed algorithms outperform the state-of-the-art techniques.

The Shepard diagrams contrasting the embeddings delivered by RMDS and HQMMDS1 are illustrated in Figure 4. HQMMDS1 embedding was obtained by the Welsch M-estimator for $a = 60$, $\lambda_1 = 3.58$ and $\lambda_2 = 100$. The number of estimated outliers for these algorithms is $\hat{S}_{RMDS} = 500$ and $\hat{S}_{HQMMDS1} = 499$. The majority of the pairwise distances $d_{ij}(\mathbf{X})$ for HQMMDS1 that were deemed as outliers lie below the diagonal $\delta_{ij} = d_{ij}$ as in RMDS. In any case, both RMDS and HQMMDS1 generate embeddings in such a way that the resulting pairwise distances are densely congregated around the diagonal line

$\delta_{ij} = d_{ij}$.

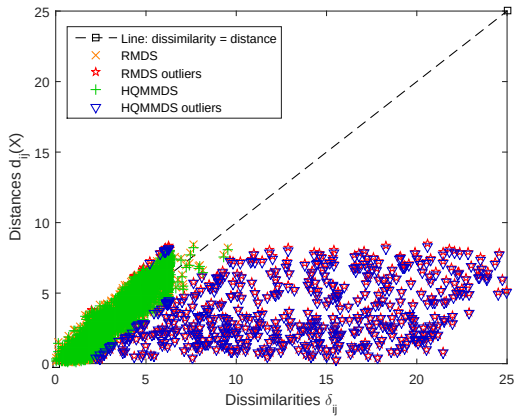
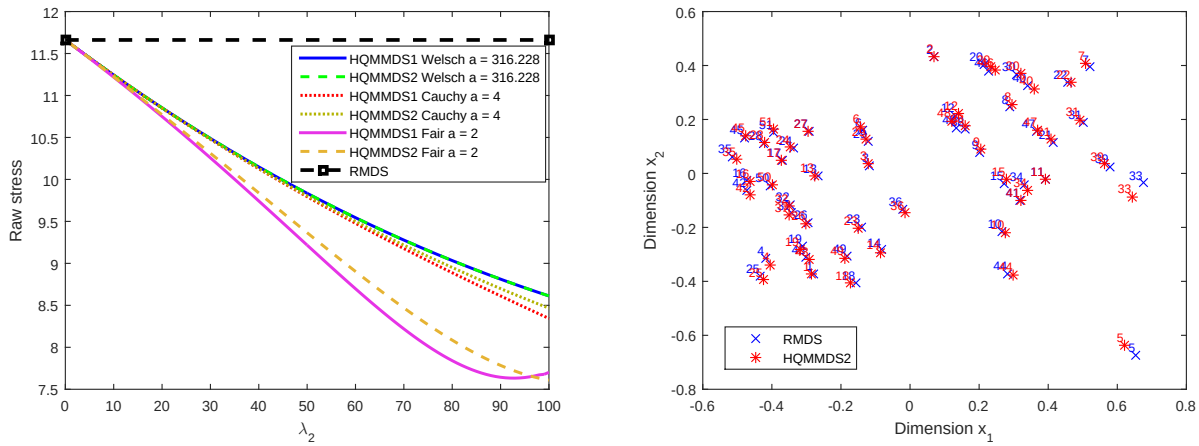


Fig. 4: Shepard diagrams for RMDS and HQMMDS1

Matching the embedding delivered by HQMMDS1 with that of REE and RMDS via Procrustes analysis reveals that the embeddings derived by HQMMDS1 and REE differ significantly, while those derived by HQMMDS1 and RMDS are approximately the same, even though the former algorithm exhibits a smaller normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ and a smaller raw stress $\sigma_r(\hat{\mathbf{X}})$ than the latter one.

C. Third Data Set

The third data set comprises real data from average Scholastic Aptitude Test (SAT) scores for the $N = 51$ states in the US, including 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 [38]. To normalize the initial values in the range between 0 and 1, the minimum value of each attribute was subtracted from the initial values of the corresponding attribute and the resulting value was divided by a measure of dispersion, such as the range (i.e., the difference between the maximum and the minimum of each attribute). Then, the dissimilarity matrix was computed according to (1). The data set was artificially contaminated by $128/(51 \cdot 50/2) = 10.04\%$ outliers, which were drawn from a uniform distribution in $[\max(\delta_{ij}), 4\max(\delta_{ij})]$. The outliers indices were chosen randomly. λ_1 was set to 0.75 in order to identify $\hat{S} = 128$ outliers with RMDS. In HQMMDS1 and HQMMDS2, the parameter a was set to 316.228, 4, and 2 for the Welsch, Cauchy, and Fair M-estimators, respectively.



(a) HQMMDS1 and HQMMDS2 raw stress $\sigma_r(\hat{\mathbf{X}})$

(b) RMDS and HQMMDS2 embeddings

Fig. 5: HQMMDS1 and HQMMDS2 embedding quality on the third set for $\varpi = 10.04\%$.

TABLE IV: Figures of merit for the embedding quality obtained by SMACOF, REE, and RMDS applied to the 3rd data set.

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

The embedding quality delivered by SMACOF, REE, and RMDS is summarized in Table IV. The reported figures of REE were measured after 8000 iterations. Due to space limitations and taking into account that the multiplicative form was proven to be much faster than the additive one, only $\sigma_r(\mathbf{X})$ for HQMMDS1 and HQMMDS2 is plotted in Figure 5a for $\lambda_2 \in [1, 100]$. The plot of $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$, in both versions, is roughly the same with that of $\sigma_r(\hat{\mathbf{X}})$ and is always smaller than that of RMDS for $\lambda_2 \in [1, 100]$. The estimated number of outliers $\hat{S}$ was proven to be rather constant and specifically for $\lambda_2 \in [1, 100]$ it admits values between 128 and 130. It is obvious that HQMMDS1 and HQMMDS2 outperform RMDS for a wide range of values admitted by λ_2.

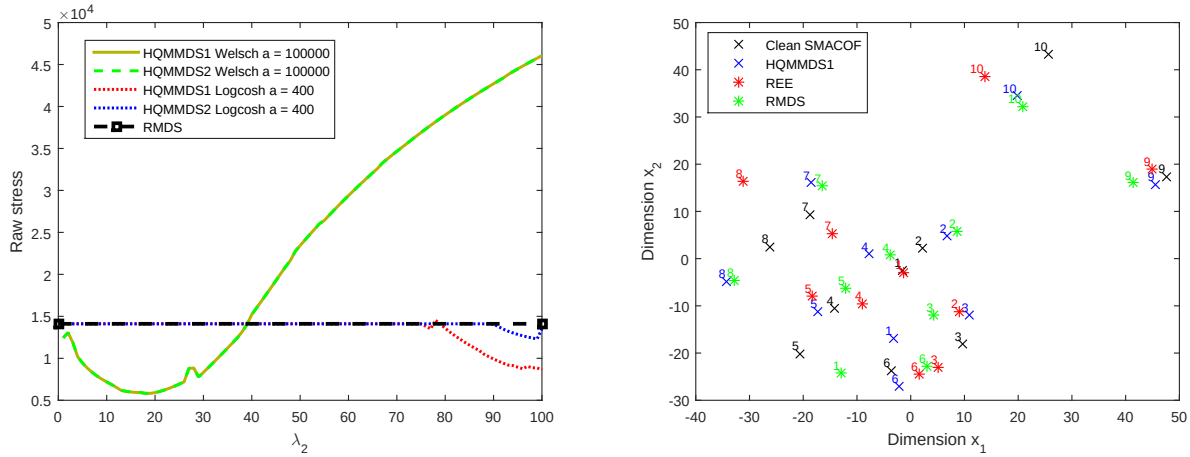
The embeddings delivered by RMDS and HQMMDS2 are shown in Figure 5b. HQMMDS2 embedding was obtained by the Fair M-estimator with $\lambda_1 = 0.75$, $\lambda_2 = 100$ and $a = 2$. It is obvious that RMDS and HQMMDS2 embeddings approximately coincide, although HQMMDS2 exhibits a smaller raw stress

$\sigma_r(\hat{\mathbf{X}})$ and a smaller normalized outlier-free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ than RMDS. This is not the case with SMACOF and REE embeddings, which differ significantly from that of HQMMDS2, as it is conferred by Procrustes' analysis.

D. Fourth Data Set

The fourth data set is derived from a delay-based scheme, entitled "sandwich probing", which is initiated to measure packet-delay differences [39]. Each sandwich probe involves three packets sent from a fixed source, namely a small packet first sent to receiver node i followed by a large packet sent to node j and finally a small packet sent once more to node i. This network includes $N = 10$ terminal (receiver) nodes, thus there are $(10 \cdot 9)/2 = 45$ terminal pairs. Each measurement is emanated from the difference between the arrival times of the first and second small packet at their terminal node i, which is relevant to the path bandwidth shared with terminal node j [39]. The sandwich probe was implemented totally $9,567$ times encompassing, inter alia, swaps between the small and the large packet receiver nodes. The mean packet-delays τ_{ij}, representing similarities between paths, constitute the outlier free (non-contaminated) data. Their transformation into dissimilarities is implemented via $\delta_{ij} = 100 \exp(-\frac{\tau_{ij}}{1000})$ as in [11]. The same transformation was imposed on minimum and maximum packet-delays for each pair of terminals in order to acquire their largest δ_{ij}^{max} and the 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 raw stress $\sigma_r(\hat{\mathbf{X}})$ for HQMMDS1 and HQMMDS2 is plotted in Figure 6a with a being set to 10^5 and 400 for the Welsch and log-cosh M-estimators, respectively. It can be seen that $\sigma_r(\hat{\mathbf{X}})$ for Welsch M-estimator admits smaller values than RMDS for $\lambda_2 \in [1, 39]$. The same conclusions are drawn for the Cauchy and Fair M-estimators for $a = 10^5$. This small range of λ_2 values ($\lambda_2 \in [1, 39]$) can be attributed to the small data set ($N = 10$). The plots of $\hat{S}$ and $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ for the Welsch M-estimator for both versions of multiplicative form look similar to $\sigma_r(\hat{\mathbf{X}})$. REE is proven to be extremely efficient on this data set. The proposed algorithms HQMMDS1 and HQMMDS2 for the Welsch M-estimator obtain slightly smaller values of $\sigma_r(\hat{\mathbf{X}})$ than REE for $\lambda_2 \in [16, 20]$. Logcosh M-estimator is proven to be better than RMDS w.r.t $\sigma_r(\hat{\mathbf{X}})$ for both versions of the multiplicative form for $\lambda_2 \in [1, 100]$. For that



(a) HQMMDS1 and HQMMDS2 raw stress $\sigma_r(\hat{\mathbf{X}})$

(b) Outlier-free SMACOF, HQMMDS1, REE, and RMDS embeddings

Fig. 6: HQMMDS1 and HQMMDS2 embedding quality on the fourth set for $\varpi = 26.67\%$.

M-estimator, the estimated number of outliers $\hat{S}$, in both versions, was proven to be relatively constant, near the value of 12, while the plot of $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ was comparable with that of $\sigma_r(\hat{\mathbf{X}})$.

The embeddings delivered by SMACOF on non-contaminated data (used as a benchmark), and REE, RMDS, and HQMMDS1 on contaminated data are shown in Figure 6b. The embeddings of REE, RMDS and HQMMDS1 were matched to that of SMACOF via Procrustes analysis. HQMMDS1 embedding was derived by the Welsch M-estimator with $\lambda_1 = 29.9$, $\lambda_2 = 19$ and $a = 10^5$. It is obvious that HQMMDS1 embedding is closer to SMACOF benchmark than those achieved by REE and RMDS. By visual inspection, REE embedding is proven to be better than that of RMDS. These deductions are also confirmed by $\sigma_r(\hat{\mathbf{X}})$ values (2371.8 for SMACOF on non-contaminated data, 5823.4 for HQMMDS1, 5941.1 for REE and 14111.4 for RMDS).

E. Discussion

It is apparent that HQAMDS and HQMMDS outperform the state-of-the-art techniques for a wide range of values admitted by λ_2. Regardless of the λ_1 value, HQMMDS and HQAMDS yield a better approximation of the true configuration than RMDS for a wide range of λ_2 values. SMACOF is extremely inefficient, while REE delivers a better embedding than SMACOF, but still this is inferior to that of RMDS in most cases. In the following, we discuss several practical issues.

M-estimator selection: The efficiency of an M-estimator depends heavily on the proper selection of the parameter a. In addition, the choice of the M-estimator is influenced by the selection of parameter λ_2 within the proposed solution (9) that yields the HQAMDS and HQMMDS, as is discussed next. Numerical tests demonstrate that the Cauchy, Fair, and Welsch M-estimators yield the most stable performance (i.e., a decreasing function of stress) for a wide range of values for λ_2. Thus, they are strongly recommended compared to other M-estimators. The greatest range of λ_2 values, where the proposed HQMMDS and HQAMDS algorithms attain a smaller raw stress $\sigma_r(\hat{\mathbf{X}})$ than RMDS, is captured with the use of ℓ_2 M-estimator.

The Geman McClure estimator was extremely inefficient in both forms. The log-cosh estimator was superior than the RMDS for a wide range of λ_2 values in both versions of the multiplicative form. However, the critical tuning of the parameter a was found to be difficult enough. Its additive form should be avoided. The $\ell_1 - \ell_2$ estimator exhibits in both forms a better performance than the RMDS for a narrow range of λ_2 values compared to that achieved by the Welsch, Cauchy, and Fair M-estimators. Huber and Tukey M-estimators exhibit comparable performance with that of Welsch, Cauchy, and Fair M-estimators, but their tuning seems to be rather difficult.

Kernel size of the potential function: A well-tuned parameter a can definitely minimize the influence of outliers and noise. It can be estimated from the data or may be determined empirically. For example, the kernel size of the Welsch M-estimator in both forms can be determined by $a^2 = \frac{\|\mathbf{LX}-\mathbf{Y}\|_F^2}{2Nd}$ [34] or alternatively by applying Silverman's rule [40]. It has been attested that both rules yield similar values with a tendency to select a rather small kernel size. Let $\hat{a}$ be the kernel size estimated by either of the two rules. A rule of thumb is to set $a = \xi\hat{a}$ for $\xi \in [2, 5]$. The parameter a of the Cauchy and Fair M-estimators can similarly be set equal to the Welsch M-estimator kernel size.

The experimental results validate that when the kernel size a of the Welsch potential function becomes larger, then the region where the MSE metric is applicable expands. Under these circumstances, the performance of the Welsch M-estimator approximates that of the ℓ_2 M-estimator. These remarks were also validated for the Cauchy, Fair, Huber and Tukey M-estimators. It should be accentuated, however, that the value of the parameter a, above which the equivalence with the ℓ_2 M-estimator takes place, is different for each M-estimator and depends highly on the data. On the contrary, a small kernel size

of the Welsch potential function shrinks the region where the ℓ_2 norm is applied, while the ℓ_1 and ℓ_0 regions are enlarged. Nonetheless, a large value of a impedes the derivation of the optimal embedding, which takes place for a larger value of λ_2. Thus, if user's objective is a wide range of λ_2 values where the proposed algorithms are more efficient than the RMDS w.r.t the raw stress $\sigma_r(\hat{\mathbf{X}})$, then a large value of a (much larger than the values predicted in [34] and [40]) should be chosen. If the objective is to find the true configuration quickly, then a small value of a is recommended.

Parameter selection within HQAMDS and HQMMDS: The performance of the proposed multiplicative forms depends on three parameters: the regularization weights λ_1 and λ_2 as well as the parameter a for each M-estimator. The performance of the proposed additive form depends also on the constant c, appearing in the minimizer function $\delta_A(\cdot)$. The choice of these parameters should be made in the following order: λ_1, a, λ_2 for the multiplicative form and λ_1, c, a, λ_2 for the additive form. Regarding the parameter c, the typical choice is $c = \phi''(0)$.

Assuming that the MAD of the nominal errors σ_ϵ is known, then λ_1 can be estimated as $\lambda_1 = 3,99\sigma_\epsilon$ borrowing the expression, which is valid for the Huber M-estimator. Otherwise, one may exploit the plot of $\hat{S}$ versus λ_1 in the implementation of RMDS. The value of λ_1 where this curve exhibits an elbow should be selected. It has been proven that the resulting embedding, for this value of λ_1, is in close proximity with that corresponding to the RMDS minimum raw stress $\sigma_r(\hat{\mathbf{X}})$.

Regarding the parameter λ_2, one may exploit the procedure for estimating $\hat{\lambda}_{CLS}$ in [41, eq. 11]. Then, $\lambda_2 = \hat{\lambda}_{CLS} \times \frac{\max(\tilde{\delta}_{ij})}{\max(\delta_{ij})}$, where $\max(\tilde{\delta}_{ij})$ is the maximum value of the contaminated dissimilarity matrix and $\max(\delta_{ij})$ is the maximum value of the initial non-contaminated dissimilarity matrix.

Algorithm comparison: Even though the additive and the multiplicative forms solve the same HQ optimization problem, their performance appears to be rather different. Theoretically speaking, both forms should yield indistinguishable results with respect to all figures of merit for the same M-estimator, provided that the parameter a for the potential function is effectively tuned for each data set. Nevertheless, the multiplicative form appears to be more adaptable, since the tuning of a is found to be simpler than that in the additive form.

The computational complexity of both forms is the same. A thorough exploration in a variety of data sets has indicated that the multiplicative form requires fewer iterations than the additive one to converge.

Hence, the multiplicative form of the HQ minimization is recommended for configurations contaminated with outliers. Furthermore, exhaustive experiments have demonstrated that the HQMMDS2 requires more iterations than the HQMMDS1 to converge, rendering it eventually more time consuming.

It has been demonstrated, in practice, that both versions of the multiplicative form for the same parameter a yield similar results with respect to all figures of merit, provided that the parameter a admits such a large value (much larger than that determined in [34] and [40]) so that the Welsch (Cauchy or Fair) M-estimator approximates the ℓ_2 estimator. This can be attributed to the fact that both methods, even though they require different iteration numbers, converge to weights that are equal to 1, delivering eventually almost identical configurations. However, when a is small, it appears that HQMMDS2 is affected by λ_2 less than HQMMDS1. In such a case, the derivation of the optimal embedding in HQMMDS2 takes place for a larger value of λ_2 than that required for HQMMDS1. When a is much smaller than the value predicted in [34] and [40], HQMMDS1 is not recommended due to a poentially unstable performance. In such a case, HQMMDS2 is preferable. On the contrary, if a is much larger than the values determined in [34] and [40], then HQMMDS1 is recommended.

To sum up, parameter a can be easily tuned within the multiplicative form of HQ. For this reason, it is more preferable than the additive form. Furthermore, it is advised to select a large value of a, to achieve stability, and then to implement HQMMDS1 since it yields similar results with the HQMMDS2, but requires less iterations to converge.

Computational time: The proposed algorithms entail fewer iterations than RMDS until convergence in many cases. However, each iteration of HQ minimization for MDS requires a slightly larger computational time than RMDS. This is due to the incorporation of $\mathbf{P}$ in the multiplicative form or $\mathbf{H}$ in the additive form, even though their estimation is not computationally demanding.

Computational complexity: At each iteration, the alternating minimization procedure in HQMMDS1 involves the updates of $\mathbf{O}$, $\mathbf{P}$, and $\mathbf{X}$. The update of an $N \times N$ matrix $\mathbf{O}$ entails the solution of $\frac{N(N-1)}{2}$ Lasso problems, each requiring $O(d)$ computations, leading totally to $O(N^2 d)$ operations. The updates of the $N \times N$ matrices $\mathbf{L}_+$ and $\mathbf{P}$ incur $O(N^2)$ operations at each iteration. The update of $\mathbf{X}$ in (33) encompasses 5 matrix multiplications with $O(N^3)$ total operations in the worst case. The inversion of the $N \times N$ matrix $(\mathbf{L}^T \mathbf{P}^{(t+1)} \mathbf{L} + \lambda_2 \mathbf{I})$ involves $O(N^3)$ operations in the worst case, depending on the inversion

algorithm. Thus, the total computational complexity per iteration is dominated by $O(N^3)$ operations in the worst case. The same applies for the remaining proposed algorithms.

Unavailability of the outlier-free dissimilarity matrix: In this case, only the normalized outlier free stress $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ and the number of outliers $\hat{S}$ can be used as figures of merit. In order to assess the efficiency of the proposed methods, HQMMDS or HQAMDS is implemented for a reasonable range of λ_2 values, having selected λ_1 according to the elbow rule. Then the embedding with the minimum value of $\hat{S}$ is selected. Extensive experiments have demonstrated that this embedding is very close to that corresponding to the minimum raw stress $\sigma_r(\hat{\mathbf{X}})$, which indicates that the true configuration is best approximated. If the number of outliers $\hat{S}$ is approximately constant for a wide range of λ_2 values, then the embedding with the minimum value of $\sigma(\hat{\mathbf{X}}, \hat{\mathbf{O}})$ is chosen. Even in highly contaminated environments, no matter if the initial dissimilarity matrix is available or not, the proposed HQMMDS and HQAMDS algorithms can find an embedding, whose distortion from the true configuration is quite smaller compared to the state of the art techniques.

Outlook: To determine whether a given dissimilarity matrix is contaminated with outliers, one may apply SMACOF as well as one of the proposed algorithms for $\lambda_2 = 0$ (or 1) and compare their raw stress $\sigma_r(\hat{\mathbf{X}})$ values. If the raw stress estimated by SMACOF is smaller than that estimated by the proposed algorithms, the dissimilarity matrix is not contaminated.

Several real data sets encompass inherently a small proportion of outliers due to remarkably disparate behavior of entities, which inevitably leads to a diversity of measurements. It has been proved that the proposed algorithms exhibit superior performance than SMACOF algorithm for a fairly small range of λ_2 values (e.g., $\lambda_2 \in [1, 5]$) in such data sets. This is due to the fact that the proposed algorithms are appropriate if and only if the dissimilarity matrix is contaminated with outliers. The inherent existence of outliers in such data sets is diminished quickly with small values of the regularization parameter λ_2. In an outlier-free dissimilarity matrix, the SMACOF has been proven to be better than the proposed algorithms.

VIII. Concluding remarks and future research directions

A new, efficient HQ framework has been introduced for solving the MDS problem in the presence of outliers. The proposed framework has been compared with three state-of-the-art MDS techniques (i.e., SMACOF, REE, RMDS) under the same conditions. The experimental results indicate that the HQ minimization, in either additive or multiplicative form, performs substantially better than the aforementioned competing techniques in all cases. For any given configuration contaminated with outliers, it has been demonstrated that it is possible to find an M-estimator so that the HQ framework outperforms the state of the art MDS techniques. Moreover, the HQMMDS2 algorithm appears to be suitable for parallel implementation, because each dimension of the coordinates can be computed separately. This is critical in big data problems emerging in scientific visualization and data mining as well as in real-time implementations of an iterative MDS in the context of sensor networks.

It is worth mentioning that all possible variants of the proposed models were not explored. For instance, the $\|\mathbf{X}\|_{2,1}$ norm as a regularization term could also be useful in the additive and multiplicative form. The estimation of the kernel size of any potential function within HQMMDS and HQAMDS could be another subject of future research.

References

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

[2] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, June 1964.

[3] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, March 1964.

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

[5] 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.

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

[7] B. Baingana and G. B. Giannakis, "Centrality-constrained graph embedding," in *Proc. 38th IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 26-31 May 2013, pp. 3113–3117.

[8] 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.

[9] J. A. Costa, N. Patwari, and A. O. Hero III, "Distributed weighted-multidimensional scaling for node localization in sensor networks," *ACM Trans. Sen. Netw.*, vol. 2, no. 1, pp. 39–64, Feb 2006.

[10] F. Mandanas and C. Kotropoulos, "A maximum correntropy criterion for robust multidimensional scaling," in *Proc. 40th IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, South Brisbane, Queensland, Australia, April 19-24 2015, pp. 1906–1910.

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

[12] V. E. McGee, "The multidimensional analysis of 'elastic' distances," *British Journal of Mathematical and Statistical Psychology*, vol. 19, pp. 181196, November 1966.

[13] J. W. Sammon, Jr., "A nonlinear mapping for data structure analysis," *IEEE Trans. Computers*, vol. C-18, no. 5, pp. 401409, May 1969.

[14] 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.

[15] 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.

[16] P. J. F. Groenen, R. Mathar, and W. Heiser, "The majorization approach to multidimensional scaling for Minkowski distances," *Journal of Classification*, vol. 12, no. 1, pp. 3–19, 1995.

[17] P. J. F. Groenen, W.J. Heiser, and J. J. Meulman, "Global optimization in least squares multidimensional scaling by distance smoothing," *Journal of Classification*, vol. 16, no. 2, pp. 225–254, 1999.

[18] P. J. F. Groenen and I. Borg, "The past, present, and future of multidimensional scaling," Econometric Institute Report EI 2013-07, Erasmus University Rotterdam, Erasmus School of Economics, Econometric Institute, Jan 2013.

[19] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer, 2005.

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

[21] L. Cayton and S. Dasgupta, "Robust Euclidean embedding," in *Proc. 23rd Int. Conf. Machine Learning*, June 2006, ICML '06, pp. 169–176.

[22] 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.

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

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

[25] 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.

[26] 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.

[27] 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.

[28] J. E. Dennis and R. E. Welsch, "Techniques for nonlinear least squares and robust regression," in *Proc. Statistical Computing Section, American Statistical Association*, Washington, D.C., USA, 1976, pp. 83–87.

[29] 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.

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

[31] 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.

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

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

[34] 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.

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

[36] K. Stovel, M. Savage, and P. Bearman, "Ascription into achievement: Models of career systems at Lloyds Bank, 1890-1970," *American Journal of Sociology*, vol. 102, no. 2, pp. 358–399, September 1996.

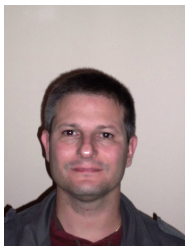
[37] A. J. Izenman, *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning*, Springer, New York, 2008.

[38] "Stats, statistical datasets," http://people.sc.fsu.edu/~jburkardt/datasets/stats/stats.html.

[39] M. Coates, R. Castro, R. Nowak, M. Gadhiok, R. King, and Y. Tsang, "Maximum likelihood network topology identification from edge-based unicast measurements," *SIGMETRICS Perform. Eval. Rev.*, vol. 30, no. 1, pp. 11–20, 2002.

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

[41] N.P. Galatsanos and A.K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Trans. Image Processing*, vol. 1, no. 3, pp. 322–336, 1992.



Fotios D. Mandanas was born in Serres, Greece in 1976. He received his B.Sc. in Telecommunications and Electronics Engineering from the Hellenic Air Force Academy in 1999, a first M.Sc. in State-of-the-Art Design and Analysis Methods in Industry (track Production Management and Industrial Administration) from the Department of Mechanical Engineering of the University of Thessaly, Volos, Greece in 2005 and a second M.Sc. in Informatics and Communications with specialization to Digital Media and Computational Intelligence from the Department of Informatics of Aristotle University of Thessaloniki, Greece in 2014. Since December 2014, he has been working toward his Ph.D. degree at the Department of Informatics, Aristotle University of Thessaloniki. His research interests include signal processing, machine learning, computational intelligence, and graph theory.

Fotios D. Mandanas is currently working in Research and Development Directorate of Hellenic Air Force Electronics Depot.

Constantine Kotropoulos (S88, M94, SM06) was born in Kavala, Greece in 1965. He received the Diploma degree with honors in Electrical Engineering in 1988 and the PhD degree in Electrical & Computer Engineering in 1993, both from the Aristotle University of Thessaloniki.

He is currently a Full Professor in the Department of Informatics at the Aristotle University of Thessaloniki. From 1989 to 1993 he was a research and teaching assistant in the Department of Electrical & Computer Engineering at the same university. In 1995, he joined the Department of Informatics at the Aristotle University of Thessaloniki as a senior researcher and served then as a Lecturer (1997-2001), an Assistant Professor (2002-2007), and an Associate Professor (2008-2015). He was a visiting research scholar in the Department of Electrical and Computer Engineering at the University of Delaware, USA during the academic year 2008-2009 and he conducted research in the Signal Processing Laboratory at Tampere University of Technology, Finland during the summer of 1993. He has co-authored 52 journal papers, 191 conference papers, and contributed 9 chapters to edited books in his areas of expertise. He is co-editor of the book Nonlinear Model-Based Image/Video Processing and Analysis (J. Wiley and Sons, 2001). His current research interests include audio, speech, and language processing; signal processing; pattern recognition; multimedia information retrieval; biometric authentication techniques, and human-centered multimodal computer interaction.

Prof. Kotropoulos was a scholar of the State Scholarship Foundation of Greece and the Bodossaki Foundation. He is a senior member of the IEEE and a member of EURASIP, IAPR, and the Technical Chamber of Greece. He is a Senior Area Editor of the IEEE Signal Processing Letters and a member of the Editorial Board of the journals: Advances in Multimedia, International Scholar Research Notices, Computer Methods in Biomechanics & Biomedical Engineering: Imaging & Visualization, and Artificial Intelligence Review. He serves as a EURASIP local liaison officer for Greece.