DISCRIMINANT ANALYSIS OF REGULARIZED MULTIDIMENSIONAL SCALING

SOHANA JAHAN
Department of Mathematics
University of Dhaka, Bangladesh
and
School of Mathematics
University of Southampton, UK

(Communicated by Wanquan Liu)

Abstract. Regularized Multidimensional Scaling with Radial Basis function (RMDS) is a nonlinear variant of classical Multi-Dimensional Scaling (cMDS). A key issue that has been addressed in RMDS is the effective selection of centers of the radial basis functions that plays a very important role in reducing the dimension preserving the structure of the data in higher dimensional space. RMDS uses data in unsupervised settings that means RMDS does not use any prior information of the dataset. This article is concerned on the supervised setting. Here we have incorporated the class information of some members of data to the RMDS model. The class separability term improved the method RMDS significantly and also outperforms other discriminant analysis methods such as Linear discriminant analysis (LDA) which is documented through numerical experiments.

1. Introduction. Nonlinear variants of classical Multi-Dimensional Scaling (MDS) involving Radial Basis Functions (RBF) was first proposed by Webb [35, 36] in the context of MDS.

Given \( N \) data points \( \{x_i\}_{i=1}^N \) in the input space \( \mathbb{R}^n \) and their associated Euclidean distances \( d_{ij} \) is defined to be \( d_{ij} = ||x_i - x_j|| \), where \( || \cdot || \) is the Euclidean norm in \( \mathbb{R}^n \). The target is to represented the data points in a lower-dimensional space \( \mathbb{R}^m \) (\( m \ll n \)) in such a way that the metric distance between \( x_i \) and \( x_j \) in the original space matches the dissimilarity \( q_{ij} \) in the lower dimensional space as closely as possible.

In [35] Webb proposed the following methodology. Firstly, the data set is mapped to another space called feature space \( \mathbb{R}^\ell \) through nonlinear function \( \Phi : \mathbb{R}^n \rightarrow \mathbb{R}^\ell \). Webb used a Radial Basis Function for this mapping: \( \Phi(x) = (\phi_1(x), \ldots, \phi_\ell(x)) \in \mathbb{R}^\ell \), with
\[
\phi_i(x) = \exp \left\{ -||x - c_i||^2/h^2 \right\}, \quad i = 1, \ldots, \ell
\]
where \( h \) is the bandwidth and \( c_i \) is the center of \( \phi_i \). Secondly, the form of data representation in \( \mathbb{R}^m \), denoted as \( f \):
\[
f(x) = W^T\Phi(x), \quad \forall \ x \in \mathbb{R}^n
\]
where $W \in \mathbb{R}^{\ell \times m}$. $f$ is a composite of a linear function (represented by the matrix $W$) and the radial basis function $\Phi$.

Thus the method seeks the best matrix $W$ that minimizes the raw STRESS (i.e., loss function):

$$\sigma^2(W) = \sum_{i,j=1}^{N} \alpha_{ij} (q_{ij}(W) - d_{ij})^2$$

where for $i,j = 1, \ldots, N$, $\alpha_{ij} > 0$ are known weights and

$$q_{ij}(W) = \|f(x_i) - f(x_j)\| = \|W^T(\Phi(x_i) - \Phi(x_j))\|.$$

The key issue in employing RBFs in MDS is to decide their centers. This includes the number of the centers to be used and then what they should be. This issue has not been well addressed in existing literature. For example, Webb [35] suggests to randomly choose the centers and then use an expensive cross-validation procedure to decide what they are. Regularized Multidimensional Scaling with Radial basis function (RMDS) takes a completely different route and regard the selection of the centers as a multi-task learning problem that has been widely studied in machine learning, see Argriou et al. [1, 2].

In RMDS [30] a regularization term $\gamma\|W\|_2^2$ is added to the stress. The optimization model thus becomes

$$\min_{W \in \mathbb{R}^{\ell \times m}} P(W) = \sigma^2(W) + \gamma\|W\|_2^2$$

here $\ell$-norm of $W$ is obtained by first computing the 2-norms of the rows $W_i$ and then the 1-norm of the vector $\|W_1\|, \|W_2\|, \ldots, \|W_\ell\|$.

$$\|W\|_{2,1} = \|W_1\| + \ldots + \|W_\ell\|,$$

here $W_i$ is the $i$th row of $W$. The $(2,1)$-norm favors a small number of nonzero rows in the matrix $W$, therefore ensuring that the common features (most effective centers) will be selected. The $(2,1)$-norm is nonsmooth (not differentiable) and the stress function $\sigma^2(W)$ is not convex. Hence, problem (6) is difficult to solve. So the majorization strategy and the techniques are nicely combined to handle the $(2,1)$-norm which led to the function

$$Q(W, C) = \sigma^2(W) + \gamma\langle WW^T, C\rangle.$$

RMDS uses data in unsupervised settings that means RMDS does not use any prior information of the dataset. This article is concerned on the supervised setting. Here we have incorporated the class information of some members of data to the RMDS model and discussed the improvement of Supervised Regularized Multidimensional Scaling (SRMDS) over RMDS. The objective function with the class separability term can be defined by :

$$J = (1 - \lambda) J_{SE} + \lambda J_{SP}$$

where $J_{SE}$ is a class separability criterion, $J_{SP}$ is a structure-preserving stress term and $\lambda (0 \leq \lambda < 1)$ determines the relative effects of these two terms. A value of $\lambda = 1.0$ gives the standard multidimensional scaling criterion with no class information. At the other extreme, $\lambda = 0$ means that emphasis is on class separability.

The rest of the article is organized as follows. In the next section, we have introduced notations and terminologies that have been used through out this article. In Section 3, we will review RMDS model which is an improvement of the RBF-MDS
model introduced by Webb [35] and On the way, we will also highlight an iterative
block-majorization method which will be applied on our proposed model. In Section
4, We will introduce the class separability term to the RMDS and study spectral
formulation for the resulting model. Numerical results on three commonly used
data sets are reported and explained in Section 5, where we demonstrate that the
supervised regularized model can significantly improve the unsupervised regularized
model as well as the original model of Webb [35]. This section also includes the
comparison between results obtained by SRMDS and other discriminant alalysis
methods such as LDA. We conclude the paper in Section 6.

2. Notations and Terminologies. Let $S^\ell$ denote the space of $\ell \times \ell$ symmetric
matrices with the standard inner product $\langle \cdot, \cdot \rangle$. Let $S^\ell_+$ denote the cone of positive
semi-definite matrices in $S^\ell$ and $S^\ell_+\!\!$ denote the set of all positive definite matrices
in $S^\ell$. Let $\mathcal{O}^\ell$ denote the set of all $\ell \times \ell$ orthonormal matrices. That is, $U \in \mathcal{O}^\ell$ if and
only if $U^TU = I$. For $C \in S^\ell_+$, we let $C^\dagger$ denote the Moore-Penrose pseudo-inverse
of $C$.

For a constant $a \in \mathbb{R}$, $a^\dagger = 1/a$ if $a \neq 0$ and $a^\dagger = 0$ otherwise. We let $\text{Tr}(C)$
denote the trace of $C$.

Suppose $C \in S^\ell_+$ has the following spectral decomposition

$$C = U \text{Diag}(\lambda_1, \ldots, \lambda_\ell) U^T,$$

where $\lambda_1 \geq \ldots \geq \lambda_\ell \geq 0$ are the eigenvalues of $C$ in nonincreasing order, $\text{Diag}(\lambda_1, \ldots, \lambda_\ell)$ is the diagonal matrix with $\lambda_i$ being on its diagonal, and $U \in \mathcal{O}^\ell$. The
pseudo-inverse of $C$ is then given by

$$C^\dagger = U \text{Diag}(\lambda_1^\dagger, \ldots, \lambda_\ell^\dagger) U^T.$$

3. Regularized Multidimensional Scaling (RMDS). In RMDS [30] a regulari-
zation term $\gamma \|W\|_{2,1}^2$ is introduced to the stress. The optimization model thus
becomes

$$\min_{W \in \mathbb{R}^{\ell \times m}} P(W) = \sigma^2(W) + \gamma \|W\|_{2,1}^2 \quad (6)$$

here $(2,1)$-norm of $W$ is obtained by first computing the 2-norms of the rows $W_i$,
and then the 1-norm of the vector $\|W_{1:1}\|, \|W_{2:1}\|, \ldots, \|W_{\ell:1}\|$.

$$\|W\|_{2,1} = \|W_{1:1}\| + \cdots + \|W_{\ell:1}\|,$$

here $W_{i:1}$ is the $i$th row of $W$. The $(2,1)$-norm favors a small number of nonzero
rows in the matrix $W$, therefore ensuring that the common features (most effective
centers) will be selected. As the $(2,1)$-norm is nonsmooth (not differentiable) and
the stress function $\sigma^2(W)$ is not convex so the problem (6) is difficult to solve.
I that case the majorization strategy and the techniques are nicely combined to
handle the $(2,1)$-norm which led us to the function

$$Q(W,C) = \sigma^2(W) + \gamma \langle WW^T, C^\dagger \rangle. \quad (7)$$

In [30], problem (6) after several modifiation is proved to be equivalent to the
problem

$$\inf \{ Q(W,D) : W \in \mathbb{R}^{\ell \times m}, D \in S^\ell_+, \text{Tr}(D) \leq 1 \}.$$

(8)
Fig. 1(a) shows the separation of the nonseparable two classes of iris data by support vector machine (SVM) algorithm. One class represented "+" and the other one is represented by "○". Over 100 runs, SRMDS model yielded at an average 1 misclassified points, while the corresponding number for RMDS was Webb’s model is 3 to 5. Fig. 1(b) shows SVM applied on 2-dimensional projection of Cancer data. Over 100 runs, SRMDS model yielded 1 to 3 misclassified points. Here support vectors in each image is bounded by "O" and misclassified points (bounded by □).

As mentioned in [30], problem (8) is not attainable, but the infimum is finite. Argyrin et. al [2] proved that such kind of problem is equivalent to the following problem, which is attainable:

\[
\min \left\{ Q(W, D) : \begin{array}{l}
W \in \mathbb{R}^{ℓ \times m} \\
D \in S_+^ℓ, \text{Tr}(D) \leq 1 \\
\text{Range}(W) \subseteq \text{Range}(D)
\end{array} \right\}.
\]

(9)

The optimal objective value of (9) equals the infimum of (8). An interesting result about (9) is that when \(W\) is fixed, minimizing \(Q(W, D)\) over \(D\) in the feasible set of (9) has a closed-form solution:

\[
D = \frac{\sqrt{WW^T}}{\text{Tr}\sqrt{WW^T}}.
\]

(10)

Here, the square root \(\sqrt{D}\) of a matrix \(D \in S_+^ℓ\) is defined to be the unique matrix \(X \in S_+^ℓ\) such that \(D = X^2\). The iterative block mejorization algorithm
Figure 2. SVM on Seeds data projected in 2 dimensional space by SRMDS is shown in these figures. Where the separation of the classes are shown using multiclass classifier.

Algorithm 3.1. Iterative Block-Majorization Method

(S.0) Initialization: Choose $W^0 \in \mathbb{R}^{\ell \times m}$ and $D^0 \in S_+^\ell$. Let $k = 0$.

(S.1) Set $V = W^k$ and update $W^k$ by

$$W^{k+1} = \arg \min_{W \in \mathbb{R}^{\ell \times m}} Q_m(W, V, D^k).$$ (11)

(S.2) Update $D^k$ by

$$D^{k+1} = \arg \min_{D \in S_+^\ell} Q(W^{k+1}, D).$$ (12)

The following remarks are useful in understanding this algorithm.
Table 1. List of datasets used in this article and their sources:

| Dataset | Dim | Class | no. of ins. | Source       |
|---------|-----|-------|-------------|--------------|
| Iris    | 4   | 3     | 150         | UCI Repository |
| Cancer  | 9   | 2     | 683         | UCI Repository |
| Seeds   | 7   | 3     | 210         | UCI Repository |

General convergence properties of Alg. 3.1 is similar to that stated as [11], which is also discussed in ([30]).

(i) Note that $D^{k+1}$ can be computed through formula (10) with $W = W^{k+1}$. The computation of $W^{k+1}$ is equivalent to solving the following equation:

\[(C + \gamma (D^k) \dagger) W = B(W^k)W^k\] (13)

with the positive semidefinite coefficient matrix $(C + \gamma (D^k) \dagger)$.

(ii) In numerical part, the algorithm is terminated whenever there was no significant change in $W$ or in $P(W)$. That is, whenever

\[\frac{\|W^{k+1} - W^k\|}{\ell^2} \leq \epsilon\]

or

\[\frac{|P(W^{k+1}) - P(W^k)|}{|P(W^k)|} \leq \epsilon\]

for a small tolerance $\epsilon > 0$, we stop the algorithm.

The selection of centres and then scaling of the data is completed in two stages.

Algorithm 3.2. Two-Stage Algorithm

S.1 Apply Alg. 3.1 to get its final iterative matrix $W \in \mathbb{R}^{\ell \times m}$. Select the most important $t_0$ centers using the method discussed in [30].

S.2 Apply the iterative block majorization algorithm of Webb [34] to minimize (2).

4. Supervised Regularized Multidimensional Scaling (SRMDS). RMDS [30] do not take any advantages of some priori information concerning the data sets. For example, some data points may be known beforehand to belong to certain class. Hence, it would be interesting to include a discriminate analysis in RMDS. Multidimensional scaling techniques to discriminant analysis have been considered by several authors [34, 16, 9, 19]. Webb [34] defined an optimization criterion that is the sum of two terms: a class separability criterion and a structure-preserving stress term which is similar to that of Koontz and Fukunaga [16]. In this work we have introduced class separability term to the objective function of Regularized Multidimensional Scaling (RMDS) and discussed the improvement of Supervised Regularized Multidimensional Scaling (SRMDS) over RMDS. The objective function with the class separability term can be defined by:

\[J = (1 - \lambda) J_{SE} + \lambda J_{SP}\]

where $J_{SE}$ is a class separability criterion, $J_{SP}$ is a structure-preserving stress term and $\lambda(0 \leq \lambda < 1)$ determines the relative effects of these two terms. A value of $\lambda = 1.0$ gives the standard multidimensional scaling criterion with no
Figure 3. Projected 2-dimensional Iris data, consisting of 3 classes. One class represented by ‘red +’ is completely separated from the other two. Training points of nonseparable two classes are represented by ‘green +’ and ‘blue +’, whereas the testing points projected by SRMDS are represented by ‘pink o’ and ‘red o’. Fig. 3(a) $\lambda = 0.1$. Fig. 3(b) $\lambda = 0.5$. Fig. 3(c) $\lambda = 0.9$, over 100 random runs. The increased value of $\lambda$ puts more weight on preservation of the structure of data.

class information. At the other extreme, $\lambda = 0$ means that emphasis is on class separability. Define the separability criterion

$$ J_{SE} = \sum_{i,j=1}^{N} \delta(i,j)\alpha_{ij}q_{ij}^2 $$

(14)

where $q_{ij}$ are the distances in the transformed space defined by equation 3. Define $\delta(i,j)$ by

$$ \delta(i,j) = \begin{cases} 
1 & \text{if } i \sim j(x_i \text{ and } x_j \text{ belongs to same class}) \\
0 & \text{otherwise,}
\end{cases} $$
and

$$\alpha_{ij} = \frac{1}{\sum_{i,j=1}^{N} \frac{1}{a_{ij}(X)}}$$  \(15\)

We define the second term \(J_{SP}\) by

$$J_{SP} = Q(W, D) = \sigma^2(W) + \gamma \langle WW^T, D^T \rangle$$

where \(\sigma^2(W) = \sum_{i,j=1}^{N} \alpha_{ij}(q_{ij}(W) - d_{ij})^2\) is identical to the loss function (2) apart from the weights \(\alpha_{ij} = \alpha_{ij}(X)\) given by equation 15. The parameter \(\lambda\) controls the relative importance of the structure preserving term to the class separability criterion. Therefore the objective function \(J\) takes the form

$$J(W) = \lambda(\sigma^2(W) + \gamma \langle WW^T, D^T \rangle) + (1 - \lambda)(\sum_{i,j=1}^{N} \delta(i, j)\alpha_{ij}q_{ij}^2)$$
for $\alpha_{ij} > 0$. This can be written as

$$J = \sum_{i,j=1}^{N} \alpha_{ij} ((1 - \lambda)\delta(i,j) + \lambda) \times \left( q_{ij} - \frac{\lambda}{(1 - \lambda)\delta(i,j) + \lambda} d_{ij}(X) \right)^2$$

$$+ \lambda \sum_{i,j=1}^{N} \alpha_{ij} \left( 1 - \frac{\lambda}{((1 - \lambda)\delta(i,j) + \lambda)^2} \right) d_{ij}^2 + \lambda \gamma (WW^T, D^T)$$

Denoting $\bar{\alpha}_{ij} = \alpha_{ij} ((1 - \lambda)\delta(i,j) + \lambda)$ and $\bar{d}_{ij} = \frac{\lambda}{(1 - \lambda)\delta(i,j) + \lambda} d_{ij}$ and ignoring the second summation (as it is independent of $q_{ij}$), we have, the minimization of $J$ is equivalent to the minimization of

$$J_e = \sum_{i,j=1}^{N} \bar{\alpha}_{ij} (q_{ij} - \bar{d}_{ij})^2 + \lambda \gamma (WW^T, D^T) \quad (16)$$

The first term of (16) is of the same form as the stress term $\sigma^2$ and thus $J_e$ has the same form of $Q(W, D)$.

Therefore $J_e$ can be minimized by Iterative block majorization algorithm 3.1 proposed in [30]. For a given $V \in \mathbb{R}^{\ell \times m}$ and $i, j = 1, \ldots, N$ define

$$B(V) = \sum_{i,j=1}^{N} c_{ij}(V)(\Phi(x_i) - \Phi(x_j))(\Phi(x_i) - \Phi(x_j))^T \in S^\ell.$$

with

$$c_{ij}(V) = \begin{cases} \lambda \alpha_{ij} d_{ij}/q_{ij}(V) & \text{if } q_{ij}(V) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and

$$C = \sum_{i,j=1}^{N} \alpha_{ij} ((1 - \lambda)\delta(i,j) + \lambda)(\Phi(x_i) - \Phi(x_j))(\Phi(x_i) - \Phi(x_j))^T.$$ 

Finally, let

$$\sigma_m^2(W, V) = \text{Tr}(WW^T CW) - 2\text{Tr}(V^TB(V)W) + \lambda \sum_{i,j=1}^{N} \alpha_{ij}d_{ij}^2.$$ 

Then, $\sigma_m^2(W, V)$ satisfies the following properties:

$$\sigma^2(W) \leq \sigma_m^2(W, V) \quad \forall W, V$$

and

$$\sigma^2(W) = \sigma_m^2(W, W).$$

Because of those properties, $\sigma_m^2(W, V)$ is called a majorization function of $\sigma^2$ at $W$. Therefore we have

$$Q_m(W, V, D) = \sigma_m^2(W, V) + \lambda \gamma (WW^T, D^T).$$

Here $Q_m$ is the majorization function of $Q(W, D)$ in the sense that

$$Q_m(W, V, D) \geq Q(W, D), \quad \forall W, V, D \quad (17)$$

and

$$Q_m(W, W, D) = Q(W, D). \quad (18)$$

Applying Alg. 3.1 we obtain the value of $W$ that minimizes $Q_m$ satisfying equation (13).
Table 2. Numerical results obtained by applying SVM on three datasets projected using discriminant analysis.

| Dataset   | MDS | RMDS | SRMDS | Improvement over RMDS |
|-----------|-----|------|-------|-----------------------|
| Iris      | 18  | 13   | 5     | 66%                   |
|           | 6   | 3    | 0     |                       |
| Cancer    | 64  | 54   | 47    | 80%                   |
|           | 9   | 5    | 1     |                       |
| Seeds C1  | 42  | 35   | 23    | 50%                   |
|           | 12  | 10   | 5     |                       |
| Seeds C2  | 20  | 16   | 10    | 33%                   |
|           | 5   | 3    | 2     |                       |
| Seeds C3  | 24  | 20   | 9     | 60%                   |
|           | 8   | 5    | 2     |                       |

Table 3. Misclassified points obtained by k-nn (3-nn) classifier on three datasets projected by SRMDS and LDA

| Dataset | LDA | SRMDS |
|---------|-----|-------|
| Iris    | 0   | 0     |
| Cancer  | 4   | 2     |
| Seeds   | 20  | 9     |

5. Numerical Experiments. We have applied our approach in several dataset. In this section we have included the results of three benchmark data sets iris, cancer and seeds from UCI repository. The details information of these datasets are given in table 2.

5.1. Parameter Setting and Performance Indicators. In the numerical experiment, the weight matrix W was initialized with random values, where $W_{ij}$ are distributed uniformly over the range [0,1]. The bandwidth parameter $h^2 = 10.0$ is taken from [34]. The tolerance $\epsilon = 10^{-4}$ is chosen for terminating the both stages of Alg. 3.2 by the rules in Remark (iii) on Alg. 3.1. The penalty parameter $\gamma$ is 1. Singular Value Decomposition is used to calculate the pseudoinverse of the matrices. The maximum number of iterations is set at $\lfloor 0.2N \rfloor$, where $N$ is the number of data samples in the data set and $\lfloor 0.2N \rfloor$ is the largest integer not greater than $0.2N$. Throughout, we set $m = 2$, which means that the original data was scaled to a data set in 2 dimensions. For each of the data sets we choose 60% random data for taining purpose and rest 40% of the data are used for testing. A random of 20% of the training data was initially selected as centers of the RBFs. Numerical experiments show that with the choices of the matrices $B$ and $C$ that incorporate class information , the projection quality improves 60 − 70% in terms of the misclassified points that we obtain applying Support Vector Machine (SVM) on the projected dataset as shown in Fig. 1(a), Fig. 1(b) and Fig. 2 and also reported in Table 2. We have compared our results with that of other discriminant analysis method such as LDA [18, 21]. In this case we have used k-nn classifier because SVM was not showing a good performance on the projected data sets obtained by LDA. Table 3 shows that SRMDS outperforms LDA in case of each dataset.

Note that the choices for $\alpha_{ij}$ and $\delta$ are quite general and the value of these parameters can be chosen in other forms [35] than those given in this section.
5.2. Choice of $\lambda$. The parameter $\lambda$ controls the relative importance of the structure preserving term to the class separability criterion. That means $\lambda(0 \leq \lambda < 1)$ determines the relative effects of these two terms. As we said earlier, a value of $\lambda = 1.0$ gives the standard multidimensional scaling criterion with no class information. At the other extreme, $\lambda = 0$ means that emphasis is on class separability. In this section we have analysed the outcomes of the algorithm for different values of $\lambda$ ($\lambda = 0.1$, $\lambda = 0.5$, $\lambda = 0.9$). Fig. 3 and Fig. 4 represents the projection of training points and testing points of iris dataset and cancer dataset respectively for different values of $\lambda$. As the value of $\lambda$ increases the algorithm preserves the structure of the data more and therefore the stress decreases which can be observed in Fig. 5.

6. Conclusion. Regularized Multidimensional Scaling with Radial basis function (RMDS) is a nonlinear variant of classical Multi-Dimensional Scaling (cMDS) where a key problem in selecting the effective centers of radial basis functions is well addressed. A new approach is taken that casts the problem as a multi-task learning problem by introducing the $(2, 1)$-norm as a regularization term to the stress function used by Webb [34]. The model RMDS hasn’t take any advantages of any priori information concerning the data sets. In this paper we have included a discriminate analysis in RMDS. That means we have incorporated the class information of some members of data to the RMDS model. Numerical experiments shows that the class separability term improved the method RMDS significantly and also outperforms other state of the art approaches.

REFERENCES

[1] A. Argyriou, T. Evgeniou and M. Pontil, Multi-task Feature Learning, in Advances in Neural Information Processing Systems (eds. B. Schoelkopf, J. Platt, and T. Hoffman), MIT Press, 2007.
[2] A. Argyriou, T. Evgeniou and M. Pontil, Convex Multi-task Feature Learning, Machine Learning, Special Issue on Inductive Transfer Learning, 73 (2008), 243–272.
[3] J. Bénasséni, Partial additive constant, J. Statist. Comput. Simul., 49 (1994), 179–193.
[4] I. Borg and P. J. F. Groenen, Modern Multidimensional Scaling. Theory and Applications, 2nd edition, Springer Series in Statistics, Springer, 2005.
[5] F. Cailliez, The analytical solution of the additive constant problem, Psychometrika, 48 (1983), 305–308.
[6] H. G. Chew and C. C. Lim, On regularisation parameter transformation of support vector machines, Journal of Industrial and Management Optimization, 5 (2009), 403–415.

[7] L. G. Cooper, A new solution to the additive constant problem in metric and multidimensional scaling, Psychometrika, 37 (1972), 311–321.

[8] T. F. Cox and M. A. Cox, Multidimensional Scaling, 2nd edition, Chapman and Hall/CRC, 2002.

[9] T. F. Cox and G. Ferry, Discriminant analysis using nonmetric multidimensional scaling, Pattern Recognition, 26 (1993), 145–153.

[10] J. de Leeuw, Applications of convex analysis to multidimensional scaling, in Recent Developments in Statistics (eds. J. Barra, F. Brodeau, G. Romier, and B. van Cutsen), North Holland Publishing Company, Amsterdam, The Netherlands, 133–145.

[11] J. de Leeuw, Block relaxation algorithms in statistics, in Information Systems and Data Analysis (eds. Bock, H.H. et al.), Springer, Berlin, (1994), 308–325.

[12] W. Glunt, T. L. Hayden, S. Hong and J. Wells, An alternating projection algorithm for computing the nearest Euclidean distance matrix, SIAM J. Matrix Anal. Appl., 11 (1990), 589–600.

[13] W. Glunt, T. L. Hayden and R. Raydan, Molecular conformations from distance matrices, J. Computational Chemistry, 14 (1993), 114–120.

[14] J. C. Gower, Some distance properties of latent root and vector methods in multivariate analysis, Biometrika, 53 (1966), 315–328.

[15] Y. Hao and F. Meng, A new method on gene selection for tissue classification, Journal of Industrial and Management Optimization, 3 (2007), 739–748.

[16] W. L. G. Koontz and K. Fukunaga, A nonlinear feature extraction algorithm using distance information, IEEE Transactions on Computers, 21 (1972), 56–63.

[17] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika, 29 (1964), 1–27.

[18] T. Li, S. Zhu and M. Ogihara, Using discriminant analysis for multi-class classification: an experimental investigation, Knowl Inf Syst., 10 (2006), 453–472.

[19] D. Lowe, Novel topographic nonlinear feature extraction using radial basis functions for concentration coding in the artificial nose, IEEE International Conference on Artificial Neural Networks, (1993), 95–99.

[20] K. V. Mardia, J. T. Kent and J. M. Bibby, Multivariate Analysis, 10th printing, Academic Press, 1995.

[21] A. M. Martinez and A. C. Kak, PCA versus LDA, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 23 (2001), 228–233.

[22] S. J. Messick and R. P Abelson, The additive constant problem in multidimensional scaling, Psychometrika, 21 (1956), 1–15.

[23] E. Pękalańska and R. P. W. Duin, The Dissimilarity Representation for Pattern Recognition: Foundations and Application, Series in Machine Perception Artificial Intelligence 64, World Scientific, 2005.

[24] H.-D. Qi, A semismooth Newton method for the nearest Euclidean distance matrix problem, SIAM Journal Matrix Analysis and Applications, 34 (2013), 67–93.

[25] H.-D. Qi and N. Xiu, A convex quadratic semidefinite programming approach to the partial additive constant problem in multidimensional scaling, Journal of Statistical Computation and Simulation, 82 (2012), 1317–1336.

[26] H.-D. Qi, N. H. Xiu and X. M. Yuan, A Lagrangian dual approach to the single source localization problem, IEEE Transactions on Signal Processing, 61 (2013), 3815–3826.

[27] H.-D. Qi and X. M. Yuan, Computing the nearest Euclidean distance matrix with low embedding dimensions, Mathematical Programming, Ser. A, 147 (2014), 351–389.

[28] K. Schittkowski, Optimal parameter selection in support vector machines, Journal of Industrial and Management Optimization, 1 (2005), 465–476.

[29] I. J. Schoenberg, Remarks to Maurice Fréchet’s article “Sur la définition axiomatique d’une classe d’espaces vectoriels distanciés applicables vectoriellement sur l’espace de Hilbert, Ann. Math., 36 (1935), 724–732.

[30] S Jahanand and H. D. Qi, Regularized multidimensional scaling with radial basis functions, Journal of Industrial and Management Optimization, 12 (2016), 543–563.

[31] S. Theodoridis and K. Koutroumbas, Pattern Recognition, Elsevier Inc., 2009.

[32] S. Theodoridis and K. Koutroumbas, An Introduction to Pattern Recognition, A MATLAB Approach, Elsevier Inc., 2010.
[33] W. S. Torgerson, *Theory and Methods for Scaling*, Wiley, New York, 1958.
[34] A. R. Webb, Multidimensional Scaling by iterative majorization using radial basis functions, *Pattern Recognition*, 28 (1995), 753–759.
[35] A. R. Webb, Nonlinear feature extraction with radial basis functions using a weighted multidimensional scaling stress measure, *Pattern Recognition*, IEEE Conference Publications, 4 (1996), 635–639.
[36] A. R. Webb, An approach to nonlinear principal component analysis using radially-symmetric kernel functions, *Statistics and Computing*, 6 (1996), 159–168.
[37] G. Young and A. S. Householder, Discussion of a set of points in terms of their mutual distances, *Psychometrika*, 3 (1938), 19–22.
[38] Y. Yuan, W. Fan and D. Pu, Spline function smooth support vector machine for classification, *Journal of Industrial and Management Optimization*, 3 (2007), 529–542.

Received October 2019; 1st revision December 2019; Final revision January 2020.

E-mail address: sjahan.mat@du.ac.bd