SUPERVISED DISTANCE PRESERVING PROJECTION USING
ALTERNATING DIRECTION METHOD OF MULTIPLIERS

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

(Communicated by Stefan Wolfgang Pickl)

ABSTRACT. Supervised Distance Preserving Projection (SDPP) is a dimension reduction method in supervised setting proposed recently by Zhu et al [43]. The method learns a linear mapping from the input space to the reduced feature space. While the method showed very promising result in regression task, for classification problems the performance is not satisfactory. The preservation of distance relation with neighborhood points forces data to project very close to one another in the projected space irrespective of their classes which ends up with low classification rate. To avoid the crowdedness of SDPP approach we have proposed a modification of SDPP which deals both regression and classification problems and significantly improves the performance of SDPP. We have incorporated the total variance of the projected co-variates to the SDPP problem which is maximized to preserve the global structure. This approach not only facilitates efficient regression like SDPP but also successfully classifies data into different classes. We have formulated the proposed optimization problem as a Semidefinite Least Square (SLS) SDPP problem. A two block Alternating Direction Method of Multipliers have been developed to learn the transformation matrix solving the SLS-SDPP which can easily handle out of sample data.

1. Introduction. In this paper we have considered a dimension reduction method in supervised setting. Supervised Distance Preserving Projection (SDPP) is a dimension reduction method proposed recently by Zhu et al [43] which showed very promising result in regression problems [20]. The basic formulation of SDPP aims to project data points in reduced feature space in such a way that the distance between data points in the projected space mimics the distance between them in the response space. The method learns a linear mapping from the input space to the reduced feature space that leads to an efficient regression design. Suppose we have \( n \) data points \( \{x_1, x_2, ..., x_n\}, x_i \in \mathbb{R}^m \) and their responses \( \{y_1, y_2, ..., y_n\}, y_i \in \mathbb{R}^k \). Assuming that the mapping \( X \rightarrow Y \) is continuous and \( X \) is well sampled, the idea is to project high dimensional data \( \{x_1, x_2, ..., x_n\} \) in a lower dimensional space \( Z \).

2010 Mathematics Subject Classification. Primary: 62H30, 68T10; Secondary: 90C25.

Key words and phrases. Supervised distance preserving projection, semidefinite programming, Alternating Direction Method of Multipliers, dimension reduction, Euclidean distance.

The research of the author was supported by the Commonwealth Scholarship Commission, UK BDCS-2012-44.
with dimensionality \( r \ll m \) by \( Z = f(X) = W^TX \) in such a way that the projection preserve distances locally between data points in the projected space (reduced feature spaces ) and the output space. In [43], SDPP seeks for the transformation matrix \( W \) that minimizes

\[
F(W) = \frac{1}{n} \sum_{i=1}^{n} \sum_{x_j \in N(x_i)} (d_{ij}(W) - \delta_{ij})^2
\]

where \( N(x_i) \) denotes a neighborhood of \( x_i \), \( d_{ij}(W) = \|z_i - z_j\|^2 = \|W^T(x_i - x_j)\|^2 \)

and \( \delta_{ij}^2 = \|y_i - y_j\|^2 \). Locality around any point \( x_i \) is controlled by its \( k \) nearest neighbors in \( N(x_i) \).

A drawback of SDPP approach is, for classification problems the preservation of local structure approach forces data of different classes to project very close to one another in the projected space which ends up with low classification rate. To avoid the crowdedness of SDPP approach we have proposed a modification of SDPP which deals both regression and classification problem and significantly improves the performance of baseline method.

Let \( \mathbb{R}^{m \times n} \) be the set of all \( m \times n \) matrices with its induced Frobenius norm \( \|\cdot\|_F \). For any \( X \in \mathbb{R}^{m \times n} \) its Frobenius norm is defined by \( \|X\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |x_{ij}|^2} \), \( S^n \) is the space of all real \( n \times n \) symmetric matrices where \( S^n_+ \) is the cone of positive semidefinite matrices in \( S^n \), \( \Pi_{S^n_+}(X) \) is the projection of a given matrix \( X \in S^n \) onto \( S^n_+ \), \( O^n \) is the set of all \( n \times n \) orthogonal matrices. For any linear operator \( \mathcal{A} : X \mapsto Y \), \( \mathcal{A}x = \langle x, \mathcal{A}y \rangle \). \( \mathcal{A}^* \) is conjugate of a linear operator \( \mathcal{A} \), where \( \mathcal{A}^*y = A_1y_1 + A_2y_2 \ldots + A_ny_n \), where \( A_i \) is the ith row of \( \mathcal{A} \), \( \langle A_x, y \rangle = \langle x, \mathcal{A}^*y \rangle \). All further notations are either standard or defined in the text of the respective sections.

In our research, we incorporated the total variance \( \sum_{i=1}^{n} \|z_i\|^2 \) of the projected co-variates to the stress \( F(W) \) SDPP problem maximize the objective function,

\[
\max \sum_{i=1}^{n} \|z_i\|^2 - \nu \sum_{i=1}^{n} \sum_{x_j \in N(x_i)} (d_{ij}^2(W) - \delta_{ij}^2)^2
\]

which can be reformulated (described in section 4) as the optimization problem:

\[
\begin{align*}
\min_{X} \quad & \langle \Psi, X \rangle - \frac{\nu}{2} \|U\|_F^2 \\
\text{s.t.} \quad & AX - U = b, \\
& X \succeq 0,
\end{align*}
\]

where \( \nu > 0 \) is the penalty parameter. To put equal emphasis on both the terms of the objective function of (1) we choose the value \( \nu = 1 \) (details discussed in sec.4).

\( X = WW^T \in \mathbb{R}^{m \times m}, \Psi = \sum_{i=1}^{n} \Psi_{ii} = \sum_{i=1}^{n} x_ix_i^T \in \mathbb{R}^{m \times m} \), \( AX \) is a vector where \( \mathcal{A}X \in \mathbb{R}^p \). Entries of \( \mathcal{A}X \) are \( \langle \Phi_{ij}, X \rangle = d_{ij}^2(W) \) and \( b = [\delta_{ij}^2] \in \mathbb{R}^p \).

The goal of our model SLS-SDPP is to determine the positive semidefinite matrix \( X \) from which the transformation matrix \( W \) can be obtained to get the projection of the data points in a lower dimensional space. The detail of this model will be discussed in section 4.

The paper is organized as follows. In Section 2, we will give a brief overview of some prominent methods of supervised dimensionality reduction. Our review of SDPP model introduced by Zhu et. al. [43] will be given in section 3. We will then introduce total variance of the projected space to the SDPP model in section 4. On the way, we will study formulation of our model as semidefinite least square.
In Section 5, we will develop a two block ADMM method for our model. We will examine the performance of the proposed method on a number of commonly used regression and classification problem in comparison with SDPP and some other state-of-the-art approaches in Section 6, where we demonstrate that the proposed model can significantly improve the original SDPP model and outperforms other models in most of the cases. We conclude the paper in Section 7.

2. Related works. Most of the research on dimension reduction are focused on unsupervised settings i.e. handles data without labels. Some recent works includes [22, 24, 27, 32, 33, 35]. Supervised dimensionality reduction methods are mostly focused on learning a subspace. Some very well known supervised DR methods includes fishers discriminant analysis (FDA), sufficient dimensionality reduction [14, 25, 42], kernel dimension reduction (KDR)[14], partial least square (PLS)[40, 41], supervised principal component analysis (SPCA) [1] and recently proposed supervised distance preserving projection method (SDPP)[43].

2.1. Fishers discriminant analysis (FDA). The most widely used supervised dimension reduction method for classification task [36, 37] is Fishers discriminant analysis (FDA) [11]and its kernelized form kernel FDA [28]. These methods maximize the ratio of between-class and within-class covariances for a good projection of data in separate classes. For a general C class problem, FDA maps the data into a (C-1) dimensional space.

2.2. Sufficient dimensionality reduction (SDR). Sufficient dimensionality reduction (SDR) [14, 25, 42] method seeks for a central subspace which is the intersection of all such subspaces containing the orthogonal transformation $U$ such that the output $Y$ and input co-variates $X$ are conditionally independent and no information about the regression is lost in reducing the dimension. But unfortunately for this approach to be successful strong assumptions have to be made on the existence of $U$. Kernel dimension reduction (KDR) is a new methodology for SDR that overcomes this problem. This method does not impose particular assumption on the underlying joint distribution of $X$ and $Y$. KDR maximizes conditional dependence by a positive definite ordering of the expected covariance operators in the probability determining reduced kernel Hilbert spaces. However KDR is computationally highly demanding.

2.3. Partial Least Square (PLS). Classical Partial Least Square (PLS) [40, 41] is a linear DR method for regression task that involves a family of techniques to analyze the relationship between blocks of data by constructing a low dimensional subspace with orthogonal latent components. At each iteration of PLS, it extracts latent vectors by maximizing the covariance between the projected co-variate and the output responses.

2.4. Supervised principal component analysis (SPCA). Supervised principal component analysis (SPCA) [1] is a generalization of Principal Component Analysis (PCA). It is based on Hilbert-Schmidt independence criterion and aims to estimate sequence of principal components that have maximal dependence on response variable. SPCA is solved by eigen decomposition of the weighted covariance matrix enhanced by the kernel of responses. Thus similar to PCA , SPCA has closed form solution and does not suffer from high computational complexity. But SPCA does
not consider local structure of data. Consequently this method cannot extract the intrinsic dimensionality of the data.

3. **Supervised distance preserving projection.** In this section we will review the SDPP model introduced by Zhu et al in [43].

3.1. **SDPP model.** The Supervised Distance Preserving Projection (SDPP) is a dimensionality reduction method that minimizes the differences between distances among projected co-variates and distances among responses locally. It also preserves the continuity of the response space. Suppose we have \( n \) data points \( \{x_1, x_2, ..., x_n\} \), \( x_i \in \mathbb{R}^m \) and their responses \( \{y_1, y_2, ..., y_n\} \).

In [43] Zhu et al. proposed the following methodology. Assuming that the mapping \( h : X \to Y \) is continuous and provided \( X \) is well sampled. It is assumed that for each point \( x \in X \) and for every \( \epsilon_y > 0 \) there exists an \( \epsilon_x > 0 \) such that \( d(x, x') < \epsilon_x \Rightarrow \delta(h(x), h(x')) < \epsilon_y \).

SDPP projects the high dimensional data \( \{x_1, x_2, ..., x_n\} \) into a lower dimensional space \( Z \) with dimensionality \( r << m \) through the linear function \( f : \mathbb{R}^m \to \mathbb{R}^r \) defined by \( f(x) = W^T x \), where the transformation matrix \( W \in \mathbb{R}^{r \times m} \). The idea of SDPP is to project the data in such a way that the local geometrical structure of the lower dimensional subspace preserves the geometrical characteristics of the response space.

Thus the method seeks for the transformation matrix \( W \) that minimizes

\[
F(W) = \frac{1}{n} \sum_{i=1}^{n} \sum_{x_j \in N(x_i)} (d_{ij}^2(W) - \delta_{ij}^2)^2
\]

where \( N(x_i) \) denotes a neighborhood of \( x_i \), \( d(, ) \) and \( \delta(, ) \) are Euclidean distances, \( d_{ij}^2(W) = \|z_i - z_j\|^2 \) and \( \delta_{ij} \) takes the form \( \delta_{ij} = \|y_i - y_j\| \) for regression task and for classification task:

\[
\delta_{ij} = \begin{cases} 
0 & \text{if } i \sim j \text{ (} x_i \text{ and } x_j \text{ belongs to same class)} \\
1 & \text{otherwise.}
\end{cases}
\]

Locality around any point \( x_i \) is controlled by its \( k \) nearest neighbors in \( N(x_i) \) where the number \( k \) is hyper-parameter of SDPP that has to be set beforehand or tuned.
from data. In [43] the value of $k$ selected by a continuity measure that is discussed briefly in 3.2.

The schematic illustration of SDPP [43] is given in Fig. 1(a). For a point $x$ in input space, consider three nearest neighbor $N(x) = \{x_1, x_2, x_3\}$. Suppose in output space the neighborhood of $y$ is $\{y_1, y_3, y_4\}$ i.e. $y_2$ is outside of the neighborhood of $y$. SDPP seeks for the transformation matrix $W$ for which $z_2 = f(x_2)$ is moved outside the neighborhood in the $Z$-space while $z_4$ is moved inside to match the local geometry in the $Y$-space as shown in Fig. 1. Thus SDPP incorporates a neighborhood graph $G_{ij}$ in the objective function defined as follows:

$$G_{ij} = \begin{cases} 1 & \text{if } i \sim j(k - N N \text{neighbor}) \\ 0 & \text{otherwise.} \end{cases}$$

Thus the objective of SDPP is to minimize

$$F(W) = \frac{1}{n} \sum_{i=1}^{n} \sum_{x_j \in N(x_i)} (d_{ij}^2(W) - \delta_{ij}^2)^2$$

which can be written equivalently as follows:

$$F(W) = \frac{1}{n} \sum_{ij} G_{ij}(d_{ij}^2(W) - \delta_{ij}^2)^2. \quad (2)$$

The rest of this research is focused on solving model (2) with the neighborhood graph $G_{ij}$.

CG method have been used in [43] to efficiently optimize the objective function of SDPP.

3.2. Continuity measure. The continuity measure, $M_{cont}^{Z \rightarrow Y}$, of the mapping $r : Z \mapsto Y$ is defined in [43, 39] by

$$M_{cont}^{Z \rightarrow Y}(k_r) = 1 - C(k_r) \sum_{i=1}^{n} \sum_{j \in V_{k_r}(i)} (r_{ij} - k)$$

where $V_{k_r}(i)$ is the set of points that are in the $k_r$-neighborhood of point $z_i$ in the projection space $Z$ but not in the response space $Y$ and $r_{ij}$ be the rank of $y_j$ in the ordering based on its distance from $y_i$. $C(k_r)$ is defined by

$$C(k_r) = \begin{cases} \frac{2n}{nk_{r}(2n-2k_{r}+1)} & \text{if } k_r < \frac{n}{2} \\ \frac{2n}{n(n-k_{r})(n-k_{r}-1)} & \text{if } k_r \geq \frac{n}{2}, \end{cases}$$

3.3. Selection of the parameter. The value of the hyper parameter $k$ can be selected in several ways as discussed in [7, 43]. I have used continuity measure discussed in section 3.2. Firstly different SDPP projection matrices $W(k)$ are learned using different locality widths $k$, in order to obtain different low-dimensional representations. Secondly, the testing input observations $X_t$ that are unseen are projected and then for each projection $Z_t(i)$ the corresponding continuity measures $M_{cont}^{Z_t \rightarrow Y}$ is calculated against the corresponding outputs $y_t$, for a sequence of region sizes $k_r$. The value of $k$ with jointly the highest continuity over the range of $k_r$ is then used to learn the final model with all the data. In [5], one of the analysis of the connectivity of nearest-neighbor graphs suggests that $k$ can be also be selected heuristically by setting $k$ to be in the order of $\log(n)$. For small sample sizes ($n \geq 100$), the value of $k$ can be selected as 10% of the available learning points ; that is $(k \approx 0.1n)$. 

4. **Formulation of SDPP as Semidefinite Least Square (SLS-SDPP).** Suppose \( \sum_{i=1}^{n} x_i = 0 \) i.e. \( \{x_i\}_{i=1}^{n} \) is already centralized. Then \( z_i = W^T x_i \), for \( i = 1, 2, \ldots, n \) is also centralized. We incorporate the total variance \( \sum_{i=1}^{n} \|z_i\|^2 = \sum_{i=1}^{n} \|W^T x_i\|^2 \) to the objective function (2). The new objective of our model is

\[
\max \sum_{i=1}^{n} \|z_i\|^2 - \frac{\nu}{n} \sum_{ij} G_{ij}(d_{ij}^2(W) - \delta_{ij}^2)^2
\]

where \( \nu > 0 \), \( G_{ij} \) is defined before in SDPP, \( (i,j) \in \xi \), \( \xi \) is the set of all possible pairs \((i,j)\) such that \( x_j \) is nearest neighbor of \( x_i \). Since for each of the \( n \) points \( x_i \), \( k \) nearest neighbors are chosen, so \( |\xi| = k \times n = p \).

The first term of (3) can be written as:

\[
\sum_{i=1}^{n} \|z_i\|^2 = \sum_{i=1}^{n} \|W^T x_i\|^2 = \sum_{i=1}^{n} \langle x_i x_i^T, WW^T \rangle = \sum_{i=1}^{n} \langle \Psi_{ii}, WW^T \rangle = \sum_{i=1}^{n} \langle \Psi_{ii}, X \rangle,
\]

where \( X = WW^T \in \mathbb{R}^{m \times m} \), \( \Psi_{ii} = x_i x_i^T \in \mathbb{R}^{m \times m} \).

Now we rewrite the square of the pairwise distance in the Z-space as

\[
d_{ij}^2(W) = \|W^T(x_i - x_j)\|^2 = \langle (x_i - x_j)(x_i - x_j)^T, WW^T \rangle
\]

Let \( \Phi_{ij} = (x_i - x_j)(x_i - x_j)^T \). Then the second term of (3) takes the form

\[
\frac{\nu}{n} \sum_{ij} G_{ij}(\langle \Phi_{ij}, X \rangle - \delta_{ij}^2)^2
\]

By definition of \( G \), \( G_{ij} = 1 \) for \( (i,j) \in \xi \). Therefore \( \sum_{ij} G_{ij}(\langle \Phi_{ij}, X \rangle - \delta_{ij}^2)^2 = \sum_{(i,j) \in \xi} (\langle \Phi_{ij}, X \rangle - \delta_{ij}^2)^2 \)

Let \( u_{ij} = (\langle \Phi_{ij}, X \rangle - \delta_{ij}^2) \). Then, the optimization model of SLS-SDPP can be written as

\[
\max \sum_{i=1}^{n} \langle \Psi_{ii}, X \rangle - \frac{\nu}{n} \sum_{ij} u_{ij}^2
\]

\[
s.t. \langle \Phi_{ij}, X \rangle - u_{ij} = \delta_{ij}
\]

\[
X \succeq 0
\]

where \( \nu > 0 \), \((i,j) \in \xi \).

Now denote \( \Psi = \sum_{i=1}^{n} \Psi_{ii} \in \mathbb{R}^{m \times m} \), \( AX = \{(\Phi_{ij}, X)\} \in \mathbb{R}^{p} \) and \( b = [\delta_{ij}^2] \in \mathbb{R}^{p}, U = [u_{ij}] \in \mathbb{R}^{p} \), such that \((i,j) \in \xi \) and \( |\xi| = p \).

We conclude this section by noting that the modified method is sensitive on the choice of the parameter \( \nu \). For example, when \( \nu << 1 \), there appeared a significant level of failure in SLS-SDPP for regression tasks. When \( \nu >> 1 \) then SLS-SDPP gives the same performance as SDPP in all the data sets (regression and classification). To get better projection using SLS-SDPP, we simply use \( \nu = 1 \) to put equal emphasis on both terms of the objective function.

Therefore the above optimization problem of SLS-SDPP takes the following form and thus our goal is to find the best matrix \( X \) which solves the following Semidefinite Least Square (SLS) problem:

\[
(P) \max \langle \Psi, X \rangle - \frac{1}{n} \|U\|^2_F \\
\text{s.t.} \ AX - U = b \\
X \in \mathbb{S}_+^m.
\]
In the next section we will study a two block ADMM method to solve SLS-SDPP problem (P).

5. Alternating Direction Method of Multipliers (ADMM) for SLS-SDPP. Alternating Direction Method of Multipliers (ADMM) is a very efficient and important algorithm that solves convex optimization problem by breaking it into smaller and easier optimizations problems. Due to its simplicity and efficiency ADMM has recently found many applications in several areas like imaging science, signal processing, machine learning etc. In this section we will study a two block ADMM to determine the optimum $X$ of (P) and therefore the best transformation matrix $W$.

Classical 2-block ADMM was first introduced by Glowinski and Marrocco [18] and Gabay and Mercier [15, 16]. Its several applications are well documented in the article of Boyd et al. [3], Eckstein and Yao [10] and in [8, 9, 19, 21, 23, 26, 29, 30, 31]. The general form of 2-block convex optimization problem is:

$$\min \{ \theta_1(z_1) + \theta_2(z_2)| \beta_1^*(z_1) + \beta_2^*(z_2) = b \}$$

where for each $i \in 1, 2, ..., q, Z_i$ is finite dimensional real Euclidean space equipped with an inner product $(\cdot, \cdot)$ and its induced norm $\| \cdot \|$. $\beta_i : X \mapsto Z_i$ is a linear map, $\beta_i^*$ is conjugate of $\beta_i$ and $b \in X$ is given. The functions $\theta_i$ are closed proper and convex and without overlapping variables.

The dual of (4) is given by

$$\max \{ \langle -b, x \rangle - \theta_1^*(S) - \theta_2^*(T) | \beta_1 x + S = 0, \beta_2 x + T = 0 \}.$$  

For given $\sigma > 0$, the augmented Lagrangian associated with (4) is given as follows

$$L_\sigma(z_1, z_2; x) := \theta_1(z_1) + \theta_2(z_2) + \langle x, \beta_1^*(z_1) + \beta_2^*(z_2) \rangle - b + \frac{\sigma}{2} \| \beta_1^*(z_1) + \beta_2^*(z_2) - b \|^2.$$  

Therefore for chosen $\tau > 0$ and $(z_1^0, z_2^0, x^0) \in \text{dom}(\theta_1) \times \text{dom}(\theta_2) \times X$ the successive iteration of classic 2-block ADMM are as follows:

$$z_1^{k+1} = \arg \min L_\sigma(z_1, z_2^k; x^k)$$

$$z_2^{k+1} = \arg \min L_\sigma(z_1^{k+1}, z_2; x^k)$$

$$x^{k+1} = x^k + \tau \sigma (\beta_1^* z_1^{k+1} + \beta_2^* z_2^{k+1} - b)$$

The convergence of 2-block ADMM is discussed in literature by several authors. For any $\tau \in (0, 2)$, convergence of 2-block ADMM has been proven first by Gabay and Mercier [16] when $\theta_1$ is strongly convex, $\beta_1^*$ is the identity mapping and $\beta_2^*$ is injective. Glowinski in [17] and Fortin and Glowinski in [12, 13] proved the convergence for $\tau \in (0, (1 + \sqrt{5})/2)$ if $\theta_2$ is general nonlinear convex function.

We note that the dual problem of (P) takes the form

$$(D) \min \ \Theta(z) = -\langle b, z \rangle + \frac{n}{4} \| z \|^2$$

$$\text{s.t. } A^* z + \Psi + S = 0$$

$$S \in S^m_+$$

For the convergence of the 2-block ADMM introduced in the next section, we need the following assumption which is a simpler version of [3, Ass-2,Sect-3.2].

**Assumption 1.** a) There exists a feasible solution $\hat{X} \in S^m_+$ of problem (P) such that $AX - U = b, \hat{X} \in \text{int}(S^m_+)$. 
b) There exists a feasible solution \( \{ \hat{S}, \hat{z} \} \in S_+^m \times \mathbb{R}^p \) of problem (D) such that 
\( A^*z + \Psi + S = 0, \hat{S} \in \text{int}(S_+^m) \).

From convex analysis [4, sec. 5.5.3] [2, Cor. 5.3.6] it is known that under assumption 1 the strong duality for (P) and (D) holds and the following Karush-Kuhn-Tucker (KKT) conditions has nonempty solution
\[
AX - U = b, \quad A^*z + \Psi + S = 0, (X, S) = 0, X \in S_+^m, S \in S_+^m. \tag{7}
\]

Now for given \( \sigma > 0 \), let \( L_\sigma(z, S; X) \), the augmented lagrange function for (D) is defined by
\[
L_\sigma(z, S; X) = -\langle b, z \rangle + \frac{n}{4} \| z \|^2 + \langle X, A^*z + \Psi + S \rangle + \frac{\sigma}{2} \| A^*z + \Psi + S \|^2
\]
\[
= -\langle b, z \rangle + \frac{n}{4} \| z \|^2 + \frac{\sigma}{2} \| A^*z + \Psi + S \|^2 + \frac{X}{\sigma} - \frac{\| X \|^2}{2\sigma}.
\]

Where \((z, S, X) \in \mathbb{R}^p \times S_+^m \times S_+^m \).

We are ready to introduce the ADMM algorithm for our SLS-SDPP problem.

**Algorithm 1. Alternating Direction Method of Multipliers**

*Given parameters \( \tau \in (0, \infty) \).*

\((S.0)\) Choose \( \sigma_0 > 0 \), \( S^0 \in S_+^m \), \( X_0 = W_0W_0^T \in S_+^m \).

Set \( z^0 = \left( \frac{\tau}{2}I + \sigma AA^* \right)^{-1} (b - AX^0 + \sigma A(-\Psi - S^0)) \). Let \( k = 0 \).

\((S.1)\) Update \( S^k \) by
\[
S^{k+1} \in \arg\min L_{\sigma_k}(z^k, S^k; X^k) = \Pi_{S_+^m} \left( -\Psi - A^*z^k - \frac{X^k}{\sigma_k} \right). \tag{8}
\]

\((S.2)\) Update \( z^k \) by
\[
z^{k+1} \in \arg\min L_{\sigma_k}(z, S^{k+1}; X^k) = \left( \frac{\tau}{2}I + \sigma AA^* \right)^{-1} (b - AX^k + \sigma A(-\Psi - S^{k+1})) \tag{9}
\]

\((S.3)\) Update \( X^k \) by
\[
X^{k+1} = X^k + \tau \sigma_k(A^*z^{k+1} + S^{k+1} + \Psi) \tag{10}
\]

\((S.4)\) Update \( \sigma_k \) by
\[
\sigma_{k+1} = \rho \sigma_k (\rho > 0) \quad \text{or} \quad \sigma_{k+1} = \sigma_k \tag{11}
\]

Note that

- In (S.1), the projection \( \Pi_{S_+^n}(X) \) of a given matrix \( X \in S^n \) onto \( S_+^n \) is the optimal solution of the problem
\[
\min \frac{1}{2} \| Y - X \|_F^2
\]
\[
\text{s.t.} \ Y \in S_+^n.
\]

Now let \( P \in O^n \) be such that \( X = P \text{Diag}(\lambda_1(X), \lambda_2(X), ..., \lambda_n(X)) P^T \).

Then \( \Pi_{S_+^n}(X) \) has the closed form \( \Pi_{S_+^n}(X) = P \text{Diag}(\max(\lambda_1(X), 0), \max(\lambda_2(X), 0), ..., \max(\lambda_n(X), 0)) P^T \).

- In (S.2), to update \( z \) we need to solve linear systems involving the operator \( AA^* \). The computation of \( AA^* \) and its (sparse) Cholesky factorization, which only needs to be done once, can be done at a moderate cost.
Table 1. List of datasets used in this article and their sources:

| Dataset                  | Dim | Class | no. of ins. | Source            |
|--------------------------|-----|-------|-------------|-------------------|
| Seismic bump             | 19  | 2     | 2584        | UCI Repository    |
| Cardiotocography         | 21  | 3     | 2126        | UCI Repository    |
| Diabetic Retinopathy     | 19  | 2     | 1115        | UCI Repository    |
| Mushroom                 | 22  | 2     | 8124        | UCI Repository    |
| Parkinson’s Telemonitoring | 16 | -     | 5875        | UCI Repository    |
| Concrete Compressive Strength | 8 | -     | 1030        | UCI Repository    |

In [34] Sun et al. discussed a similar type 3-block semi-proximal ADMM for conic optimization problem.

The convergence of Algorithm 1 for solving problem (D) follows from the following theorem established in [34].

**Theorem 1.** If the assumption 1 holds and if \( A \) is surjective. Then the sequence \((S^k, z^k, X^k)\) generated by the Algorithm 1 is well defined. Furthermore under the condition that either (a) \( \tau \in (0, 2) \) or (b) \( \tau \geq 2 \) but \( \sum_{k=0}^{\infty} \|S^{k+1} + A^* z^{k+1} + \Psi\|^2 < \infty \) the sequence \((S^k, z^k, X^k)\) converges to unique limit say \((S^\infty, z^\infty, X^\infty)\) satisfying the KKT conditions (7).

It is important to note that the algorithm does not optimizes the projection matrix \( W \) directly. It optimizes the positive semidefinite matrix \( X = WW^T \).

The projection matrix \( W \) can be computed as the square root of the matrix \( X \). If \( X \) is optimal then \( W \) being the square root of a positive semidefinite matrix \( X \) is unique and hence would be the optimal. Alternatively \( W \) can be computed applying singular value decomposition (SVD) on \( X \). In our numerical part we have applied SVD on \( X \). The \( i \)th column of \( W \) is calculated as \( \sqrt{\lambda_i} p_i \) where \( \lambda_i \) and \( p_i \) are the \( i \)th eigenvalue and eigenvector respectively. However, if \( X \) is not symmetric, there would be many such \( W \). Hence, it is reasonable to ask which one is best. But in our case \( X \) is a positive semidefinite matrix.

6. **Numerical experiments.** In this section, we study the effectiveness of the proposed SLS-SDPP in comparison with SDPP [43], SPCA [1], PLS [40, 41], KDR [14] and FDA [28] methods. The performance of the methods are evaluated on several regression and classification problems.

All tests have been carried out using the 64-bit version of MATLAB R2015a on a Windows 7 desktop with 64-bit operating system having Intel(R) Core(TM) 2 Duo CPU of 3.16GHz and 4.0GB of RAM.

For our algorithm, we measure the accuracy of an approximate optimal solution \((X, S, z)\) by using the following relative residual [34]

\[
\eta = \max\{\eta_P, \eta_D, \eta_{S^+}, \eta_C\},
\]

where \( \eta_P = \frac{\|AX - b + \frac{\chi}{2}\|}{1 + \|b\|} \), \( \eta_D = \frac{\|A^* z + S + \Psi\|}{1 + \|\Psi\|} \), \( \eta_{S^+} = \frac{\|H_{sp}(-X)\|}{1 + \|X\|} \) and \( \eta_C = \frac{\|X, S\|}{1 + \|X\| + \|S\|} \).

We compute the relative duality gap \( \eta_g = \frac{\langle \varphi, X \rangle - \langle b, z \rangle}{1 + \|\varphi\|} \) as well. We terminate the algorithm when \( \eta < 10^{-6} \).

6.1. **Parameter setting and performance indicators.** For each of the data sets, randomly 60% of the data were initially selected as training data. In the numerical experiment, the weight matrix \( W \) in \( X = WW^T \) is initialized using PCA.
on the training data set. The value of the neighborhood \( k \) is selected as \( k \leq 10 \) using the continuity measure (3.2). \( \tau \) is set to 1.618 as suggested in [34]

The maximum number of iterations is set at \([0.2N]\), where \( N \) is the number of data samples in the data set and \([0.2N]\) is the largest integer not greater than \( 0.2N \).

For regression problems, root mean squared error (RMSE) and mean absolute error (MAE) are calculated to compare the regression accuracy of our algorithm with some other methods mentioned above.

For the classification problems, 1-Nearest Neighbor rule is used on the projected low dimensional dataset to assign them into different classes. For each of the dataset, the classification error rate is calculated as the ratio of number of misclassified points to the total number of test samples.

6.2. Regression: In this section we evaluate the performance of proposed method on two real world regression problems obtained from UCI repository and compare with other methods in terms of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). We will use a number of graphs to show the improvement of SLS-SDPP over SDPP and some other leading methods.

Two very important and well known data set Parkinsons Telemonitoring (PT) and Concrete Compressive Strength (CCS) [6] Data Set are considered for experimental evaluation. Both of the datasets are preprocessed by mean centering and normalized to unit variance. After the dimension reduction step, a simple linear model is used for regression. Randomly 60% of the total data are used for training and 40% for testing. 100 such samples are evaluated and their average results are shown in Table 2 and 3.

**Table 2. Average RMSE and MAE for test set prediction of Parkinson Telemonitoring dataset**

| Method   | RMSE (mean±std) | MAE (mean±std) |
|----------|-----------------|----------------|
| SLS-SDPP | 10.6751±1.1481  | 8.3503±0.8525  |
| SDPP     | 10.7934±1.2371  | 8.7459±0.8378  |
| PLS      | 10.8133±1.2806  | 8.7822±0.8883  |
| SPCA     | 10.8006±1.2449  | 8.7714±0.8555  |
| KDR      | 10.8478±1.3032  | 8.8080±0.9139  |

**Table 3. Average RMSE and MAE (mean±std) for the test set prediction on Concrete Compressive Strength Data Set**

| Error | Dim | SLS-SDPP | SDPP | PLS | SPCA | KDR |
|-------|-----|----------|------|-----|------|-----|
| RMSE  |     | 10.4649±1.2072 | 10.5241±1.3220 | 12.7666±1.5159 | 12.8090±2.0429 | 13.7425±2.7046 |
|       |     | 10.4589±1.1536 | 10.4075±1.5903 | 11.8629±1.1837 | 12.8712±1.9074 | 11.6399±2.6641 |
|       |     | 10.4629±1.1648 | 10.4079±1.5915 | 10.9379±1.1096 | 12.8376±1.9316 | 11.5165±2.2178 |
|       |     | 10.4450±1.2848 | 10.5890±1.3001 | 10.5108±1.9431 | 12.8674±1.8535 | 10.1320±1.7599 |
|       |     | 10.2932±0.7866 | 10.7247±0.9038 | 10.4432±1.1883 | 12.9647±1.5869 | 10.2933±1.6507 |
|       |     | 10.9910±2.7090 | 10.4830±2.7228 | 10.3459±4.1480 | 10.4770±2.1142 | 10.4473±1.1485 |
|       |     | 10.4595±1.7052 | 10.5313±1.7228 | 10.4480±1.0746 | 10.4520±2.0604 | 10.4514±1.9052 |
|       |     | 10.4349±1.1134 | 10.4342±1.0034 | 10.4342±1.0034 | 10.4342±1.0034 | 16.3019±1.8078 |

| Error | Dim | SLS-SDPP | SDPP | PLS | SPCA | KDR |
|-------|-----|----------|------|-----|------|-----|
| MAE   |     | 8.5871±1.0882 | 8.5687±1.0991 | 10.3999±1.9745 | 10.4451±1.8685 | 10.5884±2.5505 |
|       |     | 8.2139±1.2042 | 8.2338±1.5291 | 9.3977±0.9294 | 10.4484±1.5633 | 9.2265±2.3954 |
|       |     | 8.2286±1.3385 | 8.2342±1.3380 | 8.3771±0.6452 | 10.4252±1.5656 | 8.9818±1.6882 |
|       |     | 8.5858±1.1262 | 8.3935±1.1461 | 8.2199±0.7943 | 10.4531±1.4777 | 6.6622±1.3565 |
|       |     | 8.6977±0.7413 | 8.5133±0.8507 | 8.1569±0.8528 | 10.5261±1.3204 | 8.0187±1.2904 |
|       |     | 8.3434±0.7236 | 8.2713±0.7528 | 8.1612±0.8389 | 8.1860±0.7925 | 8.1659±0.8225 |
|       |     | 8.2747±0.5815 | 8.2820±0.5787 | 8.1860±0.7809 | 8.1872±0.7735 | 8.1868±0.7903 |
|       |     | 8.1852±0.7413 | 8.1842±0.7477 | 8.1842±0.7477 | 8.1842±0.7477 | 13.1414±1.5359 |
Figure 2. Average Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) with error bars for prediction of test set of Parkinsons Telemonitoring Data Set obtained by SLS-SDPP, SDPP, PLS, SPCA and KDR. The bar diagram represents almost same performance for all the methods in terms of RMSE. In terms of MAE, SLS-SDPP outperforms all other methods.

Figure 3. Continuity measure with respect to different $k$ and $k_r$ for (a) Parkinsons Telemonitoring Data: Figure suggests to choose the neighborhood size $k = 8$ since highest continuity measure is obtained at $k = 8$ (b) Concrete Compressive Strength Data: Highest continuity measure is obtained at $k = 10$ therefore $k = 10$ is chosen as the neighborhood size.
In [38], Parkinson’s Telemonitoring data set is verified to be well fitted at 6 dimensional space. So here we used our approach to obtain the best 6 relevant features. The value of the parameter \(k\) is chosen to be 8 using continuity measure shown in Fig. 3(a). Table 2 and Table 3 reports the regression accuracy of the data sets for each of the four methods in terms of average root mean square error (RMSE) and mean absolute error (MAE) for each of the five methods. The red colored value indicates the minimum error in fitting the data and blue colored values indicate best estimation of each method and bold numbers at each row indicates lowest error along that dimension. The table shows that the data sets are best fitted by SLS-SDPP compared to other 4 methods. Small values of std indicates the stability of our algorithm. By SLS-SDPP method CCS data was best fitted at dimension 5. KDR showed the same performance in terms of MAE which can also be verified from Fig. 4. SDPP achieved its best estimation at lowest dimension \(D = 2\) which is useful for visualization purpose.

### 6.3. Classification:

This section is focused on classification problems. A number of real world benchmarking data set from UCI machine learning repository (http://archive.ics.uci.edu/ml/datasets.html) are considered listed in Table 1 to illustrate the performance of SLS-SDPP compared to other supervised dimensionality reduction methods SDPP, SPCA and KDR in classification task. In addition, we also compared our algorithm with Fishers Discriminant Analysis (FDA). For each of the dataset 60% of the data is used to calculate the transformation matrix which is used to predict the class of remaining 40% data using nearest-neighbor classifier.

Fig. 5 depicts error rate of testing data along different projection dimension.

For CTG data, SLS-SDPP projected the data with lowest error rate 0.1940 at dimension 3 and the error rate remained much lower than all other methods in the following dimensions. SDPP produces its best estimation at 5 dimensional space with error rate 0.2661 and for SPCA and KDR the minimum error rate is 0.2515.
Figure 5. Classification error rates for different projection dimension computed by algorithm SLS-SDPP, SDPP, SPCA, KDR and FDA. (a) CTG data (b) Seismic bump data (c) Diabetic Retinopathy data (d) Mushroom data. Figures illustrate that best performance is achieved by SLS-SDPP. The error rate for this method remained consistently lower than other methods.

Table 4. Average error rate of class prediction of test set for CTG data

| Error Rate | Dim | SLS-SDPP | SDPP | SPCA | KDR | FDA |
|------------|-----|----------|------|------|-----|-----|
| 2          | 0.2251 | 0.2865 | 0.2739 | 0.2593 | 0.208 |
| 3          | 0.1940 | 0.2827 | 0.2661 | 0.2992 | 0.208 |
| 4          | 0.1949 | 0.2943 | 0.3314 | 0.2427 | 0.208 |
| 5          | 0.1969 | 0.2661 | 0.3372 | 0.2437 | 0.208 |
| 6          | 0.1969 | 0.2749 | 0.2710 | 0.2115 | 0.208 |
| 7          | 0.1988 | 0.2827 | 0.2768 | 0.2193 | 0.208 |
| 8          | 0.1979 | 0.2768 | 0.2817 | 0.2300 | 0.208 |
| 9          | 0.1988 | 0.2700 | 0.2612 | 0.2315 | 0.208 |
| 10         | 0.1949 | 0.2690 | 0.2515 | 0.2412 | 0.208 |
and 0.2115 obtained at projection dimension 9 and 7 respectively which can also be observed from Table 4. Since this data set is of 3 classes, for FDA the solution rank is 2. Therefore the error rate remains constant for any dimension \( m \geq 2 \). For Seismic Bump data, lowest error rate 3.28% is obtained at dimension 1 and all the five methods were successful to achieve this accuracy. SLS-SDPP shows similar performance in case of CTG and Diabetic Retinopathy data. From Table 6 it can be observed that SLS-SDPP can classify the test data at dimension 2 more accurately than all other methods. The next better performance is obtained by SPCA at \( D = 4 \) as shown in Fig. 5(c) whereas FDA failed to produce a convincing projection. For Mushroom data, best estimation is obtained at \( D = 9 \) by SLS-SDPP and from \( D = 4 \) the error rate in SLS-SDPP remained consistently lower than all other methods.

**Table 5.** Average classification error rate of test set for Seismic bump data

| Error | Dim  | SLS-SDPP | SDPP  | SPCA  | KDR  | FDA  |
|-------|------|----------|-------|-------|------|------|
| 1     | 0.0328 | 0.0328 | 0.0328 | 0.0328 | 0.0328 | 0.0328 |
| 2     | 0.0701 | 0.0707 | 0.1004 | 0.0688 | 0.0328 | 0.0328 |
| 3     | 0.0701 | 0.0669 | 0.0694 | 0.0669 | 0.0328 | 0.0328 |
| 4     | 0.0701 | 0.0720 | 0.0676 | 0.0720 | 0.0328 | 0.0328 |
| 5     | 0.0701 | 0.0720 | 0.0732 | 0.0726 | 0.0328 | 0.0328 |
| 6     | 0.0701 | 0.0713 | 0.0789 | 0.0728 | 0.0328 | 0.0328 |
| 7     | 0.0701 | 0.0713 | 0.0789 | 0.0727 | 0.0328 | 0.0328 |
| 8     | 0.0701 | 0.0713 | 0.0795 | 0.0729 | 0.0328 | 0.0328 |
| 9     | 0.0701 | 0.0713 | 0.0795 | 0.0730 | 0.0328 | 0.0328 |
| 10    | 0.0701 | 0.0713 | 0.0795 | 0.0730 | 0.0328 | 0.0328 |

**Table 6.** Average error rate of class prediction of test set for Diabetic Retinopathy data

| Error | Dim  | SLS-SDPP | SDPP  | SPCA  | KDR  | FDA  |
|-------|------|----------|-------|-------|------|------|
| 1     | 0.6382 | 0.6382 | 0.6382 | 0.6382 | 0.6382 | 0.6382 |
| 2     | 0.2678 | 0.3390 | 0.2963 | 0.2934 | 0.6382 | 0.6382 |
| 3     | 0.2678 | 0.3105 | 0.3618 | 0.2963 | 0.6382 | 0.6382 |
| 4     | 0.2678 | 0.3191 | 0.2877 | 0.3048 | 0.6382 | 0.6382 |
| 5     | 0.2678 | 0.2934 | 0.2906 | 0.3134 | 0.6382 | 0.6382 |
| 6     | 0.2678 | 0.2849 | 0.2877 | 0.3048 | 0.6382 | 0.6382 |
| 7     | 0.2735 | 0.3191 | 0.2963 | 0.3048 | 0.6382 | 0.6382 |
| 8     | 0.2735 | 0.3191 | 0.2934 | 0.3134 | 0.6382 | 0.6382 |
| 9     | 0.2735 | 0.2906 | 0.2906 | 0.3134 | 0.6382 | 0.6382 |
| 10    | 0.2678 | 0.3020 | 0.3077 | 0.3048 | 0.6382 | 0.6382 |

**7. Conclusion.** This paper introduces a supervised dimension reduction method Semidefinite least square supervised distance preserving projection SLS-SDPP which is a modification of recently proposed supervised distance preserving projection (SDPP) [43]. SDPP is a supervised learning method whose basic formulation aims to preserve distances locally between data point in the projected space (reduced feature space) and the output space. For each data point, the local structure is preserved by keeping the distance of \( k \) nearest neighbors. The value of parameter \( k \)
is chosen by a continuity measure. Though the methods work very well in regression task but for classification problems the preservation of local structure approach forces data to project very close to one another in the projected space irrespective of their classes which ends up with low classification rate.

To avoid the crowdedness of SDPP approach we have proposed a modification of SDPP which deals both regression and classification problem and significantly improves the performance of SDPP. In our research, we have incorporated the total variance of the projected co-variates to the SDPP problem which prevents data of different classes to stay close therefore preserves the global structure. Thus the purpose of our model is to keep the distance relation with neighbors (local structure) and at the same time to preserve the global structure by maximizing the total variance. This approach not only facilitates efficient regression like SDPP but also successfully classifies data into different classes.

We have reformulated modified SDPP as a semidefinite least square (SLS-SDPP) model. A two block Alternating Direction method of Multipliers (ADMM) have been developed to solve the SLS-SDPP problem. Several real world data sets are considered to demonstrate the performance of our model in reducing the dimension of data by comparing the results with the baseline method and some other state-of-art approaches Supervised Principal Component Analysis (SPCA), Partial Least Square (PLS), Kernel Dimension Reduction (KDR) and Fishers Discriminant Analysis (FDA). Experimental evaluation shows that, in most of the cases our method successfully projects the data into lower dimensional space that preserves the most effective features. In regression task the proposed method showed an equivalent or superior performance than all other methods. In classification problems, SLS-SDP significantly improves SDPP and outperforms some other existing state-of-the-art-approaches.

REFERENCES

[1] E. Barshan, A. Ghodsi, Z. Azimifar and M. Z. Jahromi, Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds, Pattern Recognit, 44 (2010), 1357–1371.
[2] J. Borwein and A. S. Lewis, Convex Analysis and Non Linear Optimization: Theory and Examples, Springer, New York, 2006.
[3] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Machine Learning, 3 (2010), 1–122.
[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[5] M. R. Brito, E. L. Chávez, A. J. Quiroz and J. E. Yukich, Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection, *Stat. Probabil. Lett.*, 35 (1997), 33–42.

[6] I. Cheng Yeh, Modeling of strength of high performance concrete using artificial neural networks, *Cement and Concrete Research*, 28 (1998), 1797–1808.

[7] F. Coronaa, Z. Zhu, A. H. d. Souza Jr, M. Mulased, E. Muruf, L. Sassuf, G. Barretob and R. Baratti, Supervised Distance Preserving Projections: Applications in the quantitative analysis of diesel fuels and light cycle oils from NIR spectra, *Journal of Process Control*, 30 (2015), 10–21.

[8] J. Eckstein and D. P. Bertsekas, On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators, *Mathematical Programming*, 55 (1992), 293–318.

[9] J. Eckstein and M. Fukushima, Some reformulations and applications of the alternating direction method of multipliers, *Large Scale Optimization: State of the Art* (1993), 115–134.

[10] J. Eckstein and W. Yao, Understanding the convergence of Alternating Direction Method of Multipliers, Theoretical and Computational Perspectives, RUTCOR Research Report, 2014.

[11] R. A. Fisher, The use of multiple measurements in taxonomic problems, *Annals of Eugenics*, 7 (1936), 179–188.

[12] M. Fortin and R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, North-Holland Publishing Co., Amsterdam, 1983.

[13] M. Fortin and R. Glowinski, *On Decomposition-Coordination Methods Using an Augmented Lagrangian*, Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, North-Holland: Amsterdam, 1983.

[14] K. Fukumizu, F. R. Bach and M. Jordan, Kernel dimension reduction in regression, *Annals of Statistics*, 37 (2009), 1871–1905.

[15] D. Gabay, Applications of the method of multipliers to variational inequalities, in Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, M. Fortin and R. Glowinski, eds. of *Studies in Mathematics*, 1989.

[16] D. Gabay and B. Mercier, A dual algorithm for the solution of nonlinear variational problems via Finite element approximation, *Computers and Mathematics with Applications*, 2 (1976), 17–40.

[17] R. Glowinski, *Lectures on Numerical Methods for Nonlinear Variational Problem*, Tata Institute of Fundamental Research Lectures on Mathematics and Physics, Tata Institute of Fundamental Research, Bombay, Notes by M. G. Vijayasundaram and M. Adimurthi, 1980.

[18] R. Glowinski and A. Marrocco, Sur l'approximation, par éléments finis d'ordre un, et la résolution, par énumération-dualité, d'une classe de problèmes de dirichlet non linéaires, *Revue Française d'Automatique, Informatique et Recherche Opératoire*, 9 (1975), 41–76.

[19] R. Glowinski and P. L. Tallec, *Augmented Lagrangian Methods for the Solution of Variational Problems*, Studies in Applied and Numerical Mathematics, 1989.

[20] J. Han, M. Kamber and J. Pei, *Data Mining: Concepts and Techniques*, Springer-Verlag, Berlin, 2011.

[21] B. He, H. Yang and S. Wang, Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities, *Journal of Optimization Theory and Applications*, 106 (2000), 337–356.

[22] S. Jahan and H. D. Qi, Regularized Multidimensional Scaling with Radial Basis Functions, *Journal of Industrial and Management Optimization*, 12 (2016), 543–563.

[23] K. Jiang, D. Sun and K.-C. Toh, Solving nuclear norm regularized and semidefinite matrix least square problems with linear equality constraints, *Discrete Geometry and Optimization*, 69 (2013), 133–162.

[24] J. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, Springer, New York, 2007.

[25] K. Li, Sliced inverse regression for dimension reduction, *J Am Stat Assoc*, 86 (1991), 316–342.

[26] X. Li, D. Sun and K.-C. Toh, A Schur complement based semi-proximal ADM for convex quadratic conic programming and extensions, *Math. Program.*, 155 (2016), Ser. A, 323–373.

[27] L. J. P. Maaten, E. Postma and H. V. D. Herik, *Dimensionality Reduction: A Comparative Review*, Technical Report TiCC-TR 2009–005, Tilburg University Technical, Tilburg, 2009.

[28] S. Mika, G. Ratsch, J. Weston, B. Scholkopf and K. Mullers, Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX, Proceedings of the IEEE Signal Processing Society Workshop*, IEEE, Piscataway, (2002), 41–48.
[29] H.-D. Qi and D. Sun, An augmented Lagrangian dual approach for the H-weighted nearest correlation matrix problem, *IMA Journal of Numerical Analysis*, 31 (2011), 491–511.

[30] 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.

[31] R. T. Rockafellar, Augmented lagrangians and applications of the proximal point algorithm in convex programming, *Mathematics of Operations Research*, 1 (1976), 97–116.

[32] S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, 290 (2000), 2323–2326.

[33] B. Schölkopf, A. Smola and K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, 10 (1998), 1299–1319.

[34] D. Sun, K.-C. Toh and L. Yang, A convergent 3-block semi-proximal alternating direction method of multipliers for conic programming with 4-type of constraints, *SIAM Journal on Optimization*, 25 (2015), 882–915.

[35] J. Tenenbaum, V. Silva and J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, 290 (2000), 2319–2323.

[36] S. Theodoridis and K. Koutroumbas, *An Introduction to Pattern Recognition, A MATLAB approach*, Elsevier Inc., 2010.

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

[38] A. Tsanas, M. A. Little, P. E. McSharry and L. O. Ramig, Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests, *IEEE Transactions on Biomedical Engineering*, 57 (2010), 884–893.

[39] J. Venna and S. Kaski, Comparison of visualization methods for an atlas of gene expression data sets, *Inf Vis*, 6 (2007), 139–154.

[40] H. Wold, Soft modeling by latent variables: The nonlinear iterative partial least squares approach, *Perspectives in Probability and Statistics, Papers in Honour of MS Bartlett*, 1975, 117–142.

[41] H. Wold, *Partial Least Squares*, Encyclopedia of Statistical Sciences, 2009.

[42] Y. Yeh Y, S. Huang, Y. Lee Nonlinear dimension reduction with kernel sliced inverse regression, *IEEE Trans Knowl Data Eng*, 21 (2009), 1590–1603.

[43] Z. Zhu, T. Similä and F. Corona, Supervised distance preserving projection, *Neural Processing Letters*, 38 (2013), 445–463.

Received February 2018; 1st revision August 2018; final revision November 2018.

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