# Robust Graph Dimensionality Reduction

**Xiaofeng Zhu, Cong Lei, Hao Yu, Yonggang Li, Jiangzhang Gan, Shichao Zhang***
Guangxi Key Lab of Multi-source Information Mining & Security,
Guangxi Normal University, Guilin, 541004, China
seanzhuxf@gmail.com, CongL_hu@163.com, yuhao.gxnu@qq.com,
stublyg@163.com, 1960412020@qq.com, zhangscgxnu@gmail.com

## Abstract

In this paper, we propose conducting Robust Graph Dimensionality Reduction (RGDR) by learning a transformation matrix to map original high-dimensional data into their low-dimensional intrinsic space without the influence of outliers. To do this, we propose simultaneously 1) adaptively learning three variables, *i.e.,* a reverse graph embedding of original data, a transformation matrix, and a graph matrix preserving the local similarity of original data in their low-dimensional intrinsic space; and 2) employing robust estimators to avoid outliers involving the processes of optimizing these three matrices. As a result, original data are cleaned by two strategies, *i.e.,* a prediction of original data based on three resulting variables and robust estimators, so that the transformation matrix can be learnt from accurately estimated intrinsic space with the helping of the reverse graph embedding and the graph matrix. Moreover, we propose a new optimization algorithm to the resulting objective function as well as theoretically prove the convergence of our optimization algorithm. Experimental results indicated that our proposed method outperformed all the comparison methods in terms of different classification tasks.

## 1 Introduction

The development of modern technology makes high-dimensional data be obtained easily. Usually, high-dimensional representation is available to accurately characterize the data and enough training samples are able to output robust models [Li *et al.*, 2015]. However, the study of high-dimensional data often has to face the issue of curse of dimensionality [Chen *et al.*, 2013], while outliers may make the constructed model deviate to the real model [Rousseeuw and Leroy, 2005; Huber, 2011]. Hence, it is very necessary to reduce the dimensions of high-dimensional data and relieve the influence of outliers for dealing with high-dimensional data [Saeys *et al.*, 2008; Nie *et al.*, 2010; Zhang *et al.*, 2006].

Based on the assumption that high-dimensional data have low-dimensional intrinsic space, dimensionality reduction has been designed to handle high-dimensional data via reducing the dimensions of original data. Common dimensionality reduction methods include feature selection and subspace learning. Feature selection is designed to find a subset of all features to best represent all the features through predefined search criteria, while subspace learning maps high-dimensional data to their low-dimensional space so that the resulting low-dimensional data can reflect essential structures of original high-dimensional data. To our knowledge, a few previous methods were designed to consider the influence of the samples, *i.e.,* ignoring the distribution of different samples, for conducting dimensionality reduction. As a result, outliers may affect the performance of dimensionality reduction models [Nie *et al.*, 2010; Zhu *et al.*, 2014a; Zhu *et al.*, 2016].

Manifold learning is one of the most popular methods of subspace learning, which usually involves two steps for conducting dimensionality reduction [Roweis and Saul, 2000]. Specifically, a graph matrix (such as a sparse $k$ Nearest Neighbor ($k$NN) graph) measuring the local or global similarity of the samples is firstly built on original high-dimensional space, and then is conducted an eigenvalue decomposition to obtain the low-dimensional subspace of original high-dimensional data. Difference among manifold learning methods lies in the construction of the graph matrix [Zhu *et al.*, 2014b; Zhu *et al.*, 2017a]. A lot of existing manifold learning methods construct the graph matrix on original high-dimensional space, which often contains redundant features and easily leads to the issue of curse of dimensionality. The constructed graph matrix is certainly inaccurate. On the other hand, the goal of these existing manifold learning methods is to preserve the local or global similarity of original high-dimensional space which cannot stand for the real similarity of the data due to the impact of outliers or redundancy [Zhu *et al.*, 2017c; Zhu *et al.*, 2018]. As a consequence, the resulting similarity is also inaccurate.

In this paper, we propose a Robust Graph Dimensionality Reduction (RGDR) method by simultaneously 1) adaptively learning three matrices, *i.e.,* the reverse graph embedding of original data, the transformation matrix mapping original

*Corresponding author: Shichao Zhang (zhangsc@mailbox.gxnu.edu.cn).

high-dimensional data into their low-dimensional space, and the graph matrix preserving the local similarity of original data in their low-dimensional intrinsic space; and 2) employing robust estimators to assign small weights to outliers and large weights to important samples so that avoiding outliers involving the processes of learning three matrices. Specifically, robust estimators assign the samples with large estimation errors (*i.e.,* outliers) small or even zero weights and the samples with small estimation errors large weights to avoid the impact of outliers. Original data are estimated by the multiplication of the reverse graph embedding and the transformation matrix under the helping of the similarity matrix, so that the similarity matrix is estimated ont the refined original data to output high-quality transformation matrix.

Compared with previous dimensionality reduction methods, our proposed method has the following contributions:

- It learns both the transformation matrix and the graph matrix on the intrinsic space which is constructed by the refined original data. It is noteworthy that previous manifold learning methods [Zhu *et al.*, 2017b; He *et al.*, 2006; Belkin and Niyogi, 2003] were proposed to construct a fixed or dynamic graph matrix to measure the similarity using original high-dimensional data to easily result in inaccurate similarity measure.
- It cleans original data through two strategies. The first cleaning is to use the reverse graph embedding and the transformation matrix to predict original data so that forming a new space where noise and redundancy are removed. The second cleaning is to employ robust estimators to relief the influence of outliers for constructing dimensionality reduction models. By contrast, a lot of existing dimensionality reduction methods do not consider to avoid the influence of outliers, and only a few of previous methods [Nie *et al.*, 2014; Li *et al.*, 2017] considered either of these two strategies for conducting dimensionality reduction [Mao *et al.*, 2015; Peng and Fan, 2017].

## 2 Approach

In this paper, we use boldface uppercase letters and boldface lowercase letters, respectively, to denote the matrices and vectors. For a matrix $\mathbf{X} = [x_{i,j}]$, its *i*-th row and *j*-th column are denoted as $\mathbf{x}^i$ and $\mathbf{x}_j$, respectively, and its Frobenius norm is denoted as $||\mathbf{X}||_F = \sqrt{\sum_i \sum_j x_{i,j}^2}$. Furthermore, we denote the transpose, the trace, and the inverse, of a matrix $\mathbf{X}$, as $\mathbf{X}^T$, $tr(\mathbf{X})$, and $\mathbf{X}^{-1}$, respectively.

### 2.1 Reverse Graph

High-dimensional data in many real-world applications have been shown to contain low-dimensional intrinsic structures [Roweis and Saul, 2000; Tenenbaum *et al.*, 2000], such as teapot image analysis [Weinberger and Saul, 2006], facial expression image analysis [Song *et al.*, 2007], and human cancer analysis [Greaves and Maley, 2012]. Specifically, existing manifold learning methods (*e.g.,* [Belkin and Niyogi, 2003; Roweis and Saul, 2000]) were designed to obtain a graph matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ from original data $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{D \times n}$ (where *D* and *n* is the number of features and samples, respectively) through two sequential steps: 1) calculating the similarity between two samples, and 2) keeping the similarity between two samples if each of them is one of *k*NNs of the other; otherwise, their similarity is set to zero. After obtaining the graph matrix, the following objective function was used to learn the transformation matrix $\mathbf{U} \in \mathbb{R}^{d \times D}$:

$$\min_{\mathbf{U}, \mathbf{U}^T\mathbf{U}=\mathbf{I}} \sum_{i,j=1}^{n} a_{i,j} \|\mathbf{U}\mathbf{x}_i - \mathbf{U}\mathbf{x}_j\|_2^2 \tag{1}$$

where $d$ ($d \ll D$) is the dimensions of intrinsic space of original high-dimensional data, $\mathbf{I}$ is an identity matrix, and $\mathbf{U}\mathbf{x}_i$ is the prediction of the *i*-th sample $\mathbf{x}_i$. Recently, instead of learning a fixed graph matrix, existing literatures were designed to simultaneously learn the graph matrix $\mathbf{S}$ and the transformation matrix $\mathbf{U}$ [Du and Shen, 2015; Nie *et al.*, 2016].

In Eq. (1), the similarity $a_{i,j}$ between the *i*-th sample $\mathbf{x}_i$ and the *j*-th sample $\mathbf{x}_j$ is learnt from original data matrix $\mathbf{X}$, which, however, often contains outliers and redundant features and thus the similarity matrix $\mathbf{A}$ may not reflect the real similarity measured on the intrinsic structure of the data. To address this issue, Mao *et al.* proposed to use reverse graph embedding to uncover the intrinsic structure of the data before dimensionality reduction [Mao *et al.*, 2015]. Specifically, they assumed that there exists a low-dimensional intrinsic space, where the new representation of $\mathbf{X}$ in the intrinsic space is denoted as $\mathbf{Z} \in \mathbb{R}^{d \times n}$ via the transformation matrix $\mathbf{W} \in \mathbb{R}^{D \times d}$. Thus Eq. (1) can be transferred to the following objective function:

$$\min_{\mathbf{W}, \mathbf{W}^T\mathbf{W}=\mathbf{I}} \sum_{i,j=1}^{n} s_{i,j} \|\mathbf{W}\mathbf{z}_i - \mathbf{W}\mathbf{z}_j\|_2^2 \tag{2}$$

There are at least two distinguished differences between Eq. (1) and Eq. (2).

Firstly, in Eq. (2), original data $\mathbf{X}$ are refined by $\mathbf{WZ}$, so $\mathbf{W}$ is the transformation matrix from $\mathbf{Z}$ to $\mathbf{WZ}$, where $\mathbf{WZ}$ is an estimation of $\mathbf{X}$. Such an estimation is available to remove the noise of $\mathbf{X}$. It is noteworthy that $\mathbf{U}$ is the transformation matrix from $\mathbf{X}$ to $\mathbf{UX}$, where the goal of $\mathbf{U}$ is to learn the new representation of $\mathbf{X}$ in their low-dimensional intrinsic space.

Secondly, the graph matrix $\mathbf{A}$ in Eq. (1) is learnt from original data, whose noise and outliers may degrade its quality. Moreover, the graph matrix $\mathbf{A}$ is fixed, *i.e.,* learning from original data independent on the learning of the transformation matrix. However, the graph matrix $\mathbf{S}$ in Eq. (2) is learnt from the intrinsic space of original data, *i.e.,* preserving the similarity of refined original data and dynamically adjusted by the update of other variables *e.g.,* $\mathbf{W}$. Even given a random initialization $\mathbf{S}$, it can be adaptively updated to its optimization during the dynamically updated processes of all the variables. By contrast, a bad $\mathbf{A}$ will affect all the results of dimensionality reduction. Obviously, the graph matrix $\mathbf{S}$ in Eq. (2) should be better than $\mathbf{A}$ in Eq. (1).

Although Eq. (2) has been shown to have distinguish advantages over previous state-of-the-art dimensionality reduction methods [Mao *et al.*, 2015]. However, the influence of outliers should be further relieved as the outliers may result in refined original data deviating to the ideal one.

## 2.2 Robust Dimensionality Reduction

Instead of learning a fixed graph matrix in previous manifold learning methods, in this paper, we propose to employ robust estimators to adaptively learning a reverse graph embedding, a transformation matrix, and a graph matrix in a uniform framework via the following objective function:

$$\begin{aligned}\min_{\mathbf{S},\mathbf{W},\mathbf{W}^T\mathbf{W}=\mathbf{I},\mathbf{Z}} & \phi_1(\|\mathbf{X}-\mathbf{W}\mathbf{Z}\|_F) \\ & + \lambda\sum_{i,j=1}^{n} s_{i,j}\phi_2(\|\mathbf{W}\mathbf{z}_i-\mathbf{W}\mathbf{z}_j\|_2) + \sigma\sum_{i=1}^{n}\|\mathbf{s}_i\|_2^2 \\ s.t. & \sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{1}=1, s_{i,i}=0, s_{i,j}\geq 0\end{aligned} \tag{3}$$

where $\phi_1$ and $\phi_2$ are predefined robust estimators [Huber, 2011], all the elements of the column vector $\mathbf{1}\in\mathbb{R}^{n\times 1}$ are 1. The constraint $\sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{1}=1$ in $\mathbf{S}$ enables to result in shift invariant similarity and the constraint $\|\mathbf{s}_i\|_2^2$ avoids the trivial solution.

In Eq. (3), the graph matrix $\mathbf{S}$ is adaptively optimized with the updated $\mathbf{W}$ and $\mathbf{Z}$, so the graph matrix is dynamically obtained. Moreover, the constraint "$\sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{1}=1, s_{i,i}=0, s_{i,j}\geq 0$" makes different rows have different number of nonzero elements, *i.e.,* different samples have different numbers of nearest neighbors.

In order to avoid the impact of outliers, the most common way is to predefine robust estimators which regard the samples with large estimation error as outliers and thus assign them small or even zero weights to reduce their influence for dimensionality reduction. Traditional robust estimators include maximum likelihood type estimator, linear combinations of order statistics, estimator based on rank transformation, repeated median, and estimator using the least median of squares, and so on [Huber, 2011; He *et al.*, 2014; Nikolova and Ng, 2005].

## 2.3 Objective Function

According to the literatures [Nikolova and Ng, 2005], if a differentiable function $\phi(t_i)$ for a constant $t_i$ satisfies the following four conditions:

$$\begin{aligned}&\phi(t_i)\geq 0;\\ &\phi(0)=0;\\ &\phi(t_i)=\phi(-t_i);\\ &\phi(t_i)\geq\phi(t_j) for |t_i|>|t_j|\end{aligned} \tag{4}$$

$t_j$ is not equivalent to $t_i$, then the optimization of the variable $\mathbf{t}=[t_1,...,t_n]$ in the above function $\phi(\mathbf{t})$ can be optimized by

$$\min_{\mathbf{t}}\ \textstyle\sum_{i=1}^{n}\phi(t_i) \Leftrightarrow \min_{\mathbf{t},\mathbf{p}}\ \sum_{i=1}^{n}\mathbf{p}t_i^2+\psi(\mathbf{p}) \tag{5}$$

where $\psi(\mathbf{p})$ is the conjugate function of $\phi(\mathbf{t})$. Specifically, the optimization of a differentiable function $\phi(\mathbf{t})$ can be transferred to adaptively optimize this variable $\mathbf{t}=[t_1,...,t_n]$ and an auxiliary variable $\mathbf{p}$ by taking the efficiency of the optimization process into account. To do this, $\psi(\mathbf{p})$ can be either an explicit function or an implicit function, as we only need to know the minimization function $\delta(\mathbf{t})$ of $\psi(\mathbf{p})$ while optimizing $\mathbf{p}$ [He *et al.*, 2014].

In this paper, for simplicity, we use the following robust estimator and the corresponding minimization function to replace both $\phi_1(x)$ and $\phi_2(x)$, *i.e.,*

$$\phi(x)=\frac{x^2}{2(1+x^2)}\ \text{ with }\ \delta(x)=\frac{1}{(1+x^2)^2}. \tag{6}$$

After replacing $\phi_1(x)$ and $\phi_2(x)$ by Eq. (6) and following Eq. (5), we transfer Eq. (3) to our final objective function as follows:

$$\begin{aligned}\min_{\mathbf{S},\mathbf{W},\mathbf{W}^T\mathbf{W}=\mathbf{I},\mathbf{Z},\mathbf{b},\mathbf{C}} & \sum_{i=1}^{n}\mathbf{b}_i\|\mathbf{x}_i-\mathbf{W}\mathbf{z}_i\|_2^2 \\ & + \lambda\sum_{i,j=1}^{n} s_{i,j}c_{i,j}\|\mathbf{W}\mathbf{z}_i-\mathbf{W}\mathbf{z}_j\|_2^2 \\ & + \psi(\mathbf{b})+\psi(\mathbf{C})+\sigma\sum_{i=1}^{n}\|\mathbf{s}_i\|_2^2 \\ s.t. & \sum_{i=1}^{n}\mathbf{s}_i^T\mathbf{1}=1, s_{i,i}=0, s_{i,j}\geq 0\end{aligned} \tag{7}$$

where $\mathbf{b}\in\mathbb{R}^{n\times 1}$ and $\mathbf{C}\in\mathbb{R}^{n\times n}$, respectively, are an auxiliary vector and an auxiliary matrix.

According to results of $\mathbf{b}$ and $\mathbf{C}$ based on literature [He *et al.*, 2011] and our objective function in Eq. (7), the values of $\mathbf{b}$ and $\mathbf{C}$, respectively, are related to the residuals of $(\mathbf{x}_i-\mathbf{W}\mathbf{z}_i)$ and $(\mathbf{W}\mathbf{z}_i-\mathbf{W}\mathbf{z}_j)$. Specifically, if the residual is large, the weight will be small or even zero. In this way, outliers (usually with large residual) can be relieved or even removed from the construction of dimensionality reduction models.

## 2.4 Convergence Analysis

By denoting the objective function value of Eq. (7) as $J(\mathbf{W},\mathbf{S},\mathbf{Z},\mathbf{b},\mathbf{C})$ and the $t$-th iteration of the variables as $\mathbf{W}^t$, $\mathbf{S}^t$, $\mathbf{Z}^t$, $\mathbf{b}^t$, $\mathbf{C}^t$, we prove the convergence of our proposed method as follows.

Based on the literature [Nikolova and Ng, 2005] and fixing $\mathbf{W}^{t+1}$, $\mathbf{S}^{t+1}$, $\mathbf{Z}^{t+1}$ and $\mathbf{C}^{t+1}$, we have

$$\begin{aligned}&J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t+1},\mathbf{b}^{t+1},\mathbf{C}^{t+1})\\ &\leq J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t+1},\mathbf{b}^{t},\mathbf{C}^{t+1})\end{aligned} \tag{8}$$

While fixing $\mathbf{W}^{t+1}$, $\mathbf{S}^{t+1}$, $\mathbf{Z}^{t+1}$ and $\mathbf{b}^t$, we have

$$\begin{aligned}&J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t+1},\mathbf{b}^{t},\mathbf{C}^{t+1})\\ &\leq J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t+1},\mathbf{b}^{t},\mathbf{C}^{t})\end{aligned} \tag{9}$$

When $\mathbf{S}^{t+1}$, $\mathbf{W}^{t+1}$, $\mathbf{b}^t$ and $\mathbf{C}^t$ are fixed, it takes a closed form solution with respect to $\mathbf{Z}$, so we have

$$\begin{aligned}&J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t+1},\mathbf{b}^{t},\mathbf{C}^{t})\\ &\leq J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t},\mathbf{b}^{t},\mathbf{C}^{t})\end{aligned} \tag{10}$$

We optimize $\mathbf{W}$ by an eigenvalue decomposition, obviously, we have

$$\begin{aligned}&J(\mathbf{W}^{t+1},\mathbf{S}^{t+1},\mathbf{Z}^{t},\mathbf{b}^{t},\mathbf{C}^{t})\\ &\leq J(\mathbf{W}^{t},\mathbf{S}^{t+1},\mathbf{Z}^{t},\mathbf{b}^{t},\mathbf{C}^{t})\end{aligned} \tag{11}$$

While fixing $\mathbf{W}^t$, $\mathbf{Z}^t$, $\mathbf{b}^t$ and $\mathbf{C}^t$, it takes a closed form solution for $\mathbf{S}$, so we have

$$\begin{aligned}&J(\mathbf{W}^{t},\mathbf{S}^{t+1},\mathbf{Z}^{t},\mathbf{b}^{t},\mathbf{C}^{t})\\ &\leq J(\mathbf{W}^{t},\mathbf{S}^{t},\mathbf{Z}^{t},\mathbf{b}^{t},\mathbf{C}^{t})\end{aligned} \tag{12}$$

| Datasets | Samples | Dimensions | Classes |
|---|---|---|---|
| Usps | 1000 | 256 | 10 |
| Lung | 73 | 325 | 7 |
| Ionosphere | 351 | 34 | 2 |
| Chess | 3196 | 36 | 2 |

Table 1: The information of used datasets.

By combining Eq. (8), Eq. (9), Eq. (10), Eq. (11) with Eq. (12), we have:

$$\begin{aligned} &J(\mathbf{W}^{t+1}, \mathbf{S}^{t+1}, \mathbf{Z}^{t+1}, \mathbf{b}^{t+1}, \mathbf{C}^{t+1}) \\ &\leq J(\mathbf{W}^{t}, \mathbf{S}^{t}, \mathbf{Z}^{t}, \mathbf{b}^{t}, \mathbf{C}^{t}) \end{aligned} \tag{13}$$

# 3 Experimental Results

We evaluated our proposed method RGDR by comparing with one baseline method and four state-of-the-art dimensionality reduction methods on four public datasets in term of two different classification tasks, *i.e.,* binary classification and multi-classification.

## 3.1 Experimental Settings

We downloaded two binary-class datasets and two multi-class benchmark datasets from public website and listed their details in Table 1.

The comparison methods included a Baseline method, two subspace learning methods, and two feature selection methods. Baseline used all features to conduct classification with Support Vector Machine (SVM). Two feature selection methods included a filter method Laplacian Score (LS) [He *et al.*, 2006], and a embedded method General Sparsity Regularized (GSR) [Peng and Fan, 2017]. Two subspace learning methods included Convex Sparse Principal Component Analysis (CSPCA) [Chang *et al.*, 2016], and dimensionality reduction via Graph Structure Learning (GSL) [Mao *et al.*, 2015].

We employed the 10-fold cross validation method to conduct experiments for all the methods. Specifically, in each experiment, we firstly used every dimensionality reduction method to reduce the dimensions of the training data, and then conducted classification using SVM on the reduced data. In each experiment, we partitioned the whole dataset into ten subsets where 9 subsets were used for training and the left one subset was used for testing. During the training process, we used a 5-fold cross validation method to conduct model selection. In model selection, we set parameters of all the comparison methods by following their corresponding literature and set the parameter $\lambda$ in our method as $\{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$, and selected the parameters' combination with the best performance for testing. We repeated each experiment ten times and reported the final results as the average of all ten times.

We evaluated all the methods using classification accuracy (ACC) for both binary classification and multi-class classification. We also employed other three evaluation metrics (such as sensitivity (SEN), specificity (SPE) and Area Under Curve (AUC)) to evaluate the performance of binary classification of all the methods.

## 3.2 Experimental Analysis

We analyzed the classification results of both multi-class classification and binary classification of all the methods. We also explained the parameters sensitivity and convergence of our proposed method on all four real datasets.

**Multi-class Classification**

We conducted dimensionality reduction on high-dimensional data to output a part of all the features (*i.e.,* 20%, 40%, 60%, and 80% of all the features) to evaluate the classification results of all the methods. We reported the resulting classification accuracy in Figure 1.

Firstly, our RGDR achieved the best classification performance, followed by GSL, GSR, CSPCA, LS, and Baseline. For instance, our method improved on average by 2.49% and 7.20%, respectively, in term of classification accuracy, compared to the best comparison method (*i.e.,* GSL) and the worst comparison method (*i.e.,* Baseline). The possible reason may be the advantages of both reverse graph embedding and robust estimators. Specifically, the methods (such as GSR and GSL) took one of them into account, so they outperformed other comparison methods. Moreover, our proposed method outperformed all the methods by considering both of these constraints. It indicated that it is reasonable for considering the reverse graph embedding from dimensionality reduction which has been concluded in [Mao *et al.*, 2015], and taking robust estimators into account for avoiding the influence of outliers for data analysis which has been demonstrated in a lot of literatures *e.g.,* [He *et al.*, 2014; Nikolova and Ng, 2005; Black and Rangarajan, 1996].

Secondly, different numbers of kept features outputted different classification performance. In some cases, the classification results of some dimensionality reduction methods were worse than Baseline which used all the features to conduct SVM classification. This is because some useful features may be removed while reducing the dimensions of high-dimensional data. However, dimensionality reduction in these cases is still necessary as reduced data may improve the computation efficiency and reduce store cost [Peng and Fan, 2017; Zhu *et al.*, 2017b; Nie *et al.*, 2014]

**Binary Classification**

We reported the classification results of binary classification of all the methods on two real datasets in Table 2.

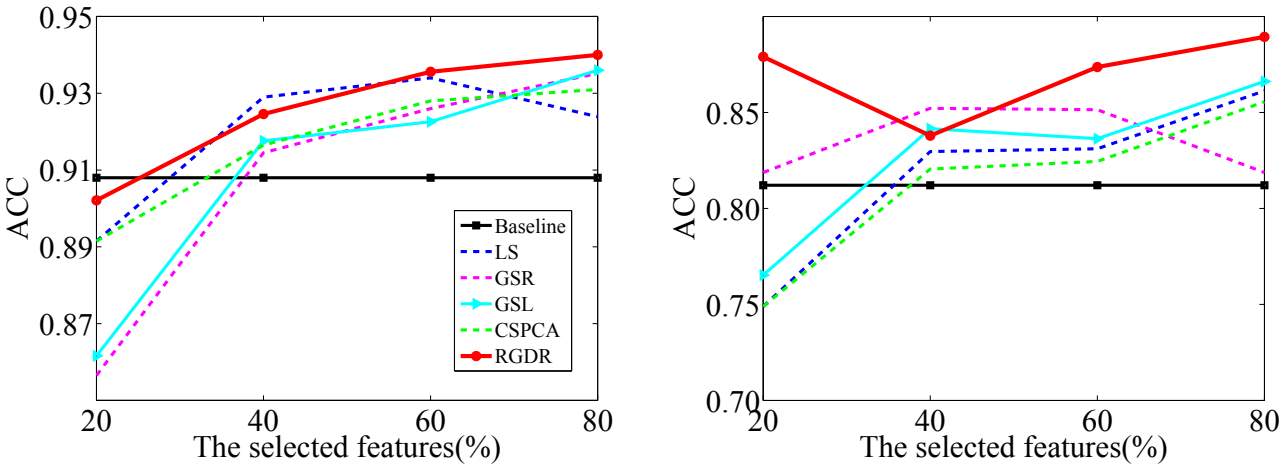
Similar to the results of multi-class classification, our proposed method still achieved the best classification performance in term of four evaluation metrics, followed by GSL, GSR, CSPCA, LS, and Baseline. More specifically, our proposed method improved on average by 2.3 %, 1.4%, 3.5%, and 2.3%, respectively, in term of classification accuracy, sensitivity, specificity, and AUC, compared to the average of all the comparison methods. This contributed to the fact that our proposed method simultaneously considered two strategies to remove noise and redundancy of original high-dimensional data, while other methods either did not take any one into account or only considered one of them. This verified again that it is reasonable to simultaneously consider the reverse graph embedding and the influence of outliers for dimensionality reduction.

**Parameters Sensitivity**

Our objective function in Eq. (7) has two parameters to be tuned, *i.e.,* $\lambda$ and $\sigma$. Based on the literature [Nie *et al.*, 2016],

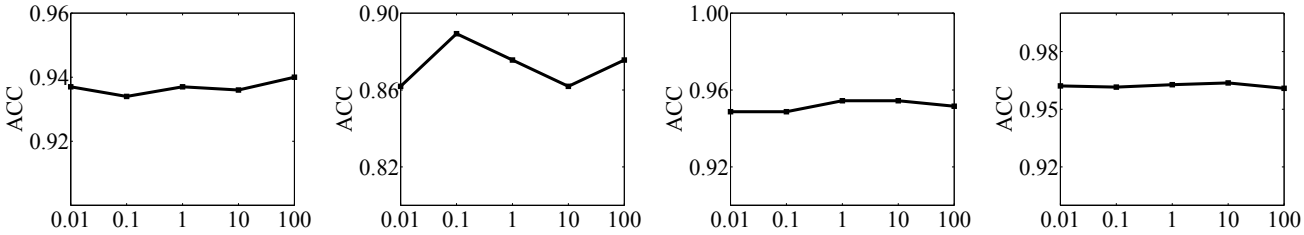
| Datasets | Ionosphere | | | | Chess | | | |
|---|---|---|---|---|---|---|---|---|
| | ACC | SEN | SPE | AUC | ACC | SEN | SPE | AUC |
| Baseline | 0.8474 | 0.9244 | 0.7378 | 0.7200 | 0.9203 | 0.9181 | 0.9227 | 0.9521 |
| LS | 0.9402 | 0.9733 | 0.8810 | 0.8967 | 0.9537 | 0.9617 | 0.9450 | 0.9913 |
| GSR | 0.9430 | 0.9778 | 0.8810 | **0.8993** | 0.9596 | 0.9599 | 0.9594 | 0.9925 |
| GSL | 0.9459 | 0.9778 | 0.8889 | 0.8514 | 0.9615 | 0.9599 | 0.9633 | 0.9930 |
| CSPCA | 0.9430 | 0.9733 | 0.8889 | 0.8851 | 0.9603 | 0.9587 | 0.9620 | 0.9931 |
| RGDR | **0.9544** | **0.9822** | **0.9048** | 0.8828 | **0.9637** | **0.9623** | **0.9653** | **0.9934** |

Table 2: Classification results of multi-class classification on two real datasets.



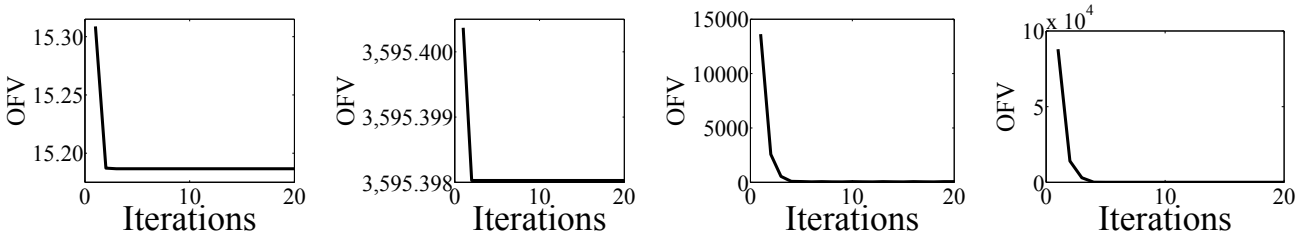

(a) Usps (b) Lung

Figure 1: Classification accuracy of all the methods on two multi-class datasets.




(a) Usps (b) Lung (c) Ionosphere (d) Chess

Figure 2: ACC results of our proposed method for different values of the parameter $\lambda$ on all datasets.




(a) Usps (b) Lung (c) Ionosphere (d) Chess

Figure 3: The variations of the Objective Function Values (OFV) of our proposed method on all the datasets.

the value of $\sigma$ can be worked out. In this section, we varied the values of $\lambda$ with the range of $\{10^{-2}, 10^{-1}, \ldots, 10^{2}\}$ to investigate the variations of the classification accuracy of our method while keeping the left features as 80% of all the features. We listed the results on four datasets in Figure 2.

From Figure 2, our method was not sensitive to the parameter setting. For example, the classification accuracy only changed about 2% on the datasets such as Usps, Ionosphere, and Chess. It indicates the robustness of our proposed method on both multi-class classification and binary classification.

**Convergence Analysis**

We proposed a new method to optimize our proposed objective function Eq. (7) and theoretically proved its convergence. We experimentally verified the convergence of objective function by investigating the variations of the objective function values of Eq. (7) at different iterations. We reported the results on all the datasets in Figure 3 while setting the stop criteria of our algorithm as $\frac{||obj(t+1)-obj(t)||_2^2}{obj(t)} \leq 10^{-5}$, where $obj(t)$ is the objective function value of Eq. (7) in the $t$-iteration.

From Figure 3, we had at least two observations: 1) the proposed algorithm sharply decreased the objective function values in the first several iterations and then began to stable; and 2) objective function converges within tens iterations on all the datasets. These conclusions indicated that our method had solved the proposed objective function in Eq. (7) and achieved fast convergence.

## 4 Conclusion

This paper has proposed a novel robust graph dimensionality reduction method using two strategies to remove the influence of noise and outliers in original high-dimensional data. Specially, the reverse graph embedding strategy makes the transformation matrix to be constructed from the low-dimensional intrinsic space, while robust estimators avoid the learning of three matrices (such as the reverse graph embedding, the transformation matrix, and the graph matrix) to be involved by the outliers. Experimental results demonstrated the effectiveness and robustness of the proposed method for two kinds of classification tasks, compared to the state-of-the-art dimensionality reduction methods.

## Acknowledgements

This work was supported in part by the China Key Research Program (Grant No: 2016YFB1000905), the Nation Natural Science Foundation of China (Grants No: 61573270, 61672177 and 81701780), the Guangxi Natural Science Foundation (Grant No: 2015GXNSFCB139011 and 2017GXNSFBA198221), the Guangxi High Institutions Program of Introducing 100 High-Level Overseas Talents, the Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing, the Guangxi Bagui Teams for Innovation and Research, the Research Fund of Guangxi Key Lab of MIMS (18-A-01-01), the PhD research startup foundation of Guangxi Normal University (Grants No: 2017BQ17), and Innovation Project of Guangxi Graduate Education under grant YCSW2018093 and YCSW2018094.

## References

[Belkin and Niyogi, 2003] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[Black and Rangarajan, 1996] Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *International Journal of Computer Vision*, 19(1):57–91, 1996.

[Chang *et al.*, 2016] Xiaojun Chang, Feiping Nie, Yi Yang, Chengqi Zhang, and Heng Huang. Convex sparse pca for unsupervised feature learning. *ACM Transactions on Knowledge Discovery from Data*, 11(1):3, 2016.

[Chen *et al.*, 2013] Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *CVPR*, pages 3025–3032, 2013.

[Du and Shen, 2015] Liang Du and Yi-Dong Shen. Unsupervised feature selection with adaptive structure learning. pages 209–218, 2015.

[Greaves and Maley, 2012] Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 2012.

[He *et al.*, 2006] Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. pages 507–514, 2006.

[He *et al.*, 2011] Ran He, Wei Shi Zheng, and Bao Gang Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 33(8):1561–1576, 2011.

[He *et al.*, 2014] Ran He, Baogang Hu, Xiaotong Yuan, and Liang Wang. M-estimators and half-quadratic minimization. In *Robust Recognition via Information Theoretic Learning*, pages 3–11. 2014.

[Huber, 2011] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. 2011.

[Li *et al.*, 2015] Zechao Li, Jing Liu, Jinhui Tang, and Hanqing Lu. Robust structured subspace learning for data representation. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):2085–2098, 2015.

[Li *et al.*, 2017] Zhihui Li, Feiping Nie, Xiaojun Chang, and Yi Yang. Beyond trace ratio: Weighted harmonic mean of trace ratios for multiclass discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2100–2110, 2017.

[Mao *et al.*, 2015] Qi Mao, Li Wang, Steve Goodison, and Yijun Sun. Dimensionality reduction via graph structure learning. In *KDD*, pages 765–774, 2015.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint l 2,1 -norms minimization. In *NIPS*, pages 1813–1821, 2010.

[Nie *et al.*, 2014] Feiping Nie, Jianjun Yuan, and Heng Huang. Optimal mean robust principal component analysis. In *ICML*, pages 1062–1070, 2014.

[Nie *et al.*, 2016] Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised feature selection with structured graph optimization. In *AAAI*, pages 1302–1308, 2016.

[Nikolova and Ng, 2005] Mila Nikolova and Michael K Ng. Analysis of half-quadratic minimization methods for signal and image recovery. *SIAM Journal on Scientific computing*, 27(3):937–966, 2005.

[Peng and Fan, 2017] Hanyang Peng and Yong Fan. A general framework for sparsity regularized feature selection via iteratively reweighted least square minimization. In *AAAI*, pages 2471–2477, 2017.

[Rousseeuw and Leroy, 2005] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 589. 2005.

[Roweis and Saul, 2000] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[Saeys *et al.*, 2008] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. *Machine learning and knowledge discovery in databases*, pages 313–325, 2008.

[Song *et al.*, 2007] Le Song, Alex Smola, Arthur Gretton, and Karsten M Borgwardt. A dependence maximization view of clustering. In *ICML*, pages 815–822, 2007.

[Tenenbaum *et al.*, 2000] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[Weinberger and Saul, 2006] Kilian Q Weinberger and Lawrence K Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *AAAI*, volume 6, pages 1683–1686, 2006.

[Zhang *et al.*, 2006] Shichao Zhang, Yongsong Qin, Xiaofeng Zhu, Jilian Zhang, and Chengqi Zhang. Optimized parameters for missing data imputation. In *PRICAI*, pages 1010–1016, 2006.

[Zhu *et al.*, 2014a] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. Multi-modality canonical feature selection for alzheimer's disease diagnosis. In *MICCAI*, pages 162–169, 2014.

[Zhu *et al.*, 2014b] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *CVPR*, pages 1542–1549, 2014.

[Zhu *et al.*, 2016] Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Dinggang Shen, and Guorong Wu. Early diagnosis of alzheimer's disease by joint feature selection and classification on temporally structured support vector machine. In *MICCAI*, pages 264–272, 2016.

[Zhu *et al.*, 2017a] Pengfei Zhu, Wencheng Zhu, Qinghua Hu, Changqing Zhang, and Wangmeng Zuo. Subspace clustering guided unsupervised feature selection. *Pattern Recognition*, 66:364–374, 2017.

[Zhu *et al.*, 2017b] Xiaofeng Zhu, Yonghua Zhu, Shichao Zhang, Rongyao Hu, and Wei He. Adaptive hypergraph learning for unsupervised feature selection. In *IJCAI*, pages 3581–3587, 2017.

[Zhu *et al.*, 2017c] Yingying Zhu, Xiaofeng Zhu, Minjeong Kim, Daniel Kaufer, and Guorong Wu. A novel dynamic hyper-graph inference framework for computer assisted diagnosis of neuro-diseases. In *IPMI*, pages 158–169. Springer, 2017.

[Zhu *et al.*, 2018] Pengfei Zhu, Qian Xu, Qinghua Hu, Changqing Zhang, and Hong Zhao. Multi-label feature selection with missing labels. *Pattern Recognition*, 74:488–502, 2018.