

Graph Scaling Cut with L1-Norm for Classification of Hyperspectral Images

Ramanarayan Mohanty
Advanced Technology
Development Centre
Indian Institute of Technology
Kharagpur, India 721302
Email: ramanarayan@iitkgp.ac.in

S L Happy
Dept. of Electrical Engineering
Indian Institute of Technology
Kharagpur, India 721302
Email: happy@iitkgp.ac.in

Aurobinda Routray
Dept. of Electrical Engineering
Indian Institute of Technology
Kharagpur, India 721302
Email: aroutray@iitkgp.ac.in

Abstract—In this paper, we propose an L1 normalized graph based dimensionality reduction method for Hyperspectral images, called as ‘L1-Scaling Cut’ (L1-SC). The underlying idea of this method is to generate the optimal projection matrix by retaining the original distribution of the data. Though L2-norm is generally preferred for computation, it is sensitive to noise and outliers. However, L1-norm is robust to them. Therefore, we obtain the optimal projection matrix by maximizing the ratio of between-class dispersion to within-class dispersion using L1-norm. Furthermore, an iterative algorithm is described to solve the optimization problem. The experimental results of the HSI classification confirm the effectiveness of the proposed L1-SC method on both noisy and noiseless data.

Index Terms—Dimensionality reduction, Hyperspectral classification, L1-norm, L1-SC, scaling cut, Supervised learning.

I. INTRODUCTION

Hyperspectral remote sensing images with high spatial and spectral resolution is used to capture the inherent properties of the surface. These hyperspectral images (HSI) contains a huge number of contiguous spectral bands. It spreads over a narrow spectral bandwidth with wealth of information content. These informations are used for characterization, identification and classification of the physical and chemical properties of the land-cover with improved accuracy. The huge spectral bands implies the high dimensional redundant HSI data. This high dimensionality is the major challenge in HSI classification. In order to overcome this challenge, dimensionality reduction (DR) is usually applied to the HSI data. An effective DR method reduces the high dimension data into low dimensional representative features. This DR method improves the classification performance by reducing computational complexity and exploring the intrinsic property of the reduced data features.

In the field of HSI processing, a large number of DR approaches have been developed during past few years. Among them in unsupervised category principal component analysis (PCA) [1] is widely used one. In addition, several supervised DR approaches have also been developed, linear discriminant analysis (LDA) [2] is the most popular classical approach. Among LDA and PCA, LDA uses the labeled information for DR and performs better than PCA in classification task. However, LDA always assumes the data distribution as Gaussian with equal variance and unimodal. Hence, it fails to handle

the real HSI data which is heteroscedastic and multimodal in nature. A graph based scaling cut (SC) [3], [4] method addresses these problem by constructing the pairwise similarity matrix among the samples of the classes.

The graph based SC method basically makes a projection of the data into a lower dimensional space by maximizing the variance of the input data points. Although, the SC method by Zhang *et al.* in [4] works well for the multimodal hyperspectral data, they are very sensitive in handling the outliers and noise in the dataset. The conventional SC works by computing the dissimilarity matrix among the data samples. This dissimilarity matrix computation is mostly done by calculating the conventional L2-norm between the samples. The square operation in L2-norm criterion magnifies the outliers [5], [6]. Therefore, the presence of outliers drift the projection vectors from the desired projection direction. Hence, dimension reduction and classification of hyperspectral data demands robust algorithms that are resistant to possible outliers.

It is found that, the L1-norm based DR method is a robust alternative to handle outliers problem in image classification [6], [7], [8], [9], [10]. Kwak *et al.* [9] computed the covariance matrix using L1-norm and proposed PCA-L1 by greedy strategy. Ke and Kanade, in [8], proposed L1-PCA by using the alternative convex method to solve the projection matrix. Similarly in [6], Wang *et al.* proposed LDA-L1 by solving the supervised LDA method using L1-norm maximization in an iterative manner. Li *et al.* [11] proposed the 2D version of the LDA (L1-2DLDA) using the L1-norm optimization.

L1-norm based LDA has achieved excellent performance for image classification [6][11]. However, our HSI data is heteroscedastic and multimodal. SC has proved its worth [4] HSI classification. Motivated by these literature, we propose a L1-norm based scaling cut method (L1-SC) for DR and classification of HSI data. In this work we formulate the SC algorithm into an L1-norm optimization problem by maximizing the ratio of between-class dissimilarity and within-class dissimilarity matrix. Then, we solve this L1-norm optimization problem by using an iterative algorithm to generate a projection matrix. The projected reduced dimension HSI data are further used for classification by using support vector machine (SVM) classifier. We analyze the classification performance by

applying it over the spectral information of two real world HSI datasets.

The rest of the paper is organized as follows. In section II, a brief introduction to the conventional L2-norm based SC method is discussed. We present the proposed L1-SC method including its objective function and algorithmic procedure for its solution in section III. Then section IV enumerates the experimental results of the proposed L1-SC method over two HSI datasets. Finally, we give the conclusive remarks to our work in section V.

II. CONVENTIONAL L2- NORM BASED GRAPH SCALING CUT CRITERION REVISITED

The purpose of SC is to determine the mapping matrix for projecting the original data into a lower dimension space. The classical LDA method is computed based on the assumption that the data distribution of each class is Gaussian with equal variance. However, the distribution of real world data is more complex than Gaussian. Hence LDA fails when data is heteroscedastic and multimodal. The major advantage of the SC over the state-of-art LDA is handling these heteroscedastic and multimodal data. This method eliminates the Gaussian distribution limitation of LDA by constructing the dissimilarity matrix among the data samples.

Let $X = (x_1, x_2, \dots, x_n) \in R^{D \times n}$ is the input training dataset, given by $\{x_i, L_i\}_{i=1}^n$. Here $L_i = \{1, 2, \dots, C\}$ is the class label of the corresponding training data with total C classes and n training data samples. The objective is to determine a projection matrix, that project the input training data of D dimensions into reduced d dimension such that $d \ll D$. The between-class dissimilarity matrix and the within-class dissimilarity matrix of SC are defined as

$$\begin{aligned} S_{B_k}^{SC} &= \sum_{x_i \in U_k} \sum_{x_j \in \bar{U}_k} \frac{1}{n_k n_{\bar{k}}} (x_i - x_j)(x_i - x_j)^T \\ S_{W_k}^{SC} &= \sum_{x_i \in U_k} \sum_{x_j \in U_k} \frac{1}{n_k n_k} (x_i - x_j)(x_i - x_j)^T \end{aligned} \quad (1)$$

where U_k represents all the samples from k th class and n_k is the total number of elements in U_k . Similarly, \bar{U}_k represents all the data points that does not belong to the k th class and $n_{\bar{k}}$ denotes the total number of elements in \bar{U}_k . $S_{B_k}^{SC}$ represents the dissimilarity between U_k class and \bar{U}_k , whereas $S_{W_k}^{SC}$ is the dissimilarity matrix within the U_k class. Based on the $S_{W_k}^{SC}$ and $S_{B_k}^{SC}$, the objective function of SC can be written as

$$\begin{aligned} Scut(W) &= \frac{\left| \sum_{k=1}^c W^T S_{B_k}^{SC} W \right|}{\left| \sum_{k=1}^c (W^T S_{W_k}^{SC} W + W^T S_{B_k}^{SC} W) \right|} \\ &= \frac{|W^T S_B^{SC} W|}{|W^T (S_W^{SC} + S_B^{SC}) W|} \\ &= \frac{|W^T S_B^{SC} W|}{|W^T S_T^{SC} W|} \end{aligned} \quad (2)$$

where $S_B^{SC} = \sum_{k=1}^c S_{B_k}^{SC}$, $S_W^{SC} = \sum_{k=1}^c S_{W_k}^{SC}$, and $S_T^{SC} = (S_B^{SC} + S_W^{SC})$ is the total dissimilarity matrix and W is the projection matrix. The dissimilarity matrix is scaled according to the size of the class. Hence, this graph cut is termed as scaling cut.

III. PROPOSED L1-NORM BASED SCALING CUT CRITERION

The conventional L2-norm based graph scaling cut criterion basically determines the projection matrix by maximizing the between-class distances, and minimizing the within-class distances to enhance the compactness among the data points. These models characterize the geometric structure of the data by computing the L2-norm. These L2-norm is computed by using square euclidean distance, which is sensitive to outliers and noises [6], [7]. These outlier elements drift the projection vectors from the desired projection directions. Hence, it reduces the flexibility of L2-norm based algorithms. To handle this issue, L1-norm based technique is widely used as a robust alternative of conventional L2-norm based technique. Motivated by the idea of L1-norm based modeling, we propose to model the graph based scaling cut criterion by using the L1-norm optimization instead of L2-norm optimization. This L1-norm based SC method is solved by following iterative algorithm

A. L1-norm based Graph Scaling Cut (L1-SC)

Inspired by the existing literatures on L1-norm based method [7], [11], we propose to maximize the SC criterion using L1-norm rather than L2-norm. The equation (2) can be simplified to a trace ratio [12] problem, which can further be reduced to the Frobenius norm, given by,

$$\begin{aligned} W^* &= \max_{W^T W = I} \frac{Tr(W^T S_B^{SC} W)}{Tr(W^T S_T^{SC} W)} \\ &= \max_{W^T W = I} \frac{\sum_{k: x_i \in U_k; x_j \in \bar{U}_k} \frac{1}{n_k n_{\bar{k}}} Tr(W^T (x_i - x_j)(x_i - x_j)^T W)}{\sum_{k: x_i \in U_k; x_j \in U_k} \frac{1}{n_k n_k} Tr(W^T (x_i - x_j)(x_i - x_j)^T W)} \\ &= \max_{W^T W = I} \frac{\sum_{k=1}^c \sum_{x_i \in U_k} \sum_{x_j \in \bar{U}_k} \frac{1}{n_k n_{\bar{k}}} \|W^T (x_i - x_j)\|_F^2}{\sum_{k=1}^c \sum_{x_i \in U_k} \sum_{x_j \in U_k} \frac{1}{n_k n_k} \|W^T (x_i - x_j)\|_F^2} \end{aligned} \quad (3)$$

As can be observed, the above objective is based on Frobenius norm, which also involves the square operations and in term, it is sensitive to outliers to noise and outliers similar to L2-norm. In order to reduce the sensitivity, we use the objective function in terms of the L1-norm. The proposed model of the objective function for L1-norm SC is defined as,

$$\begin{aligned}
v_{opt} &= \max_{v^T v=1} \frac{\sum_{k=1}^c \sum_{x_i \in U_k} \sum_{x_j \in \bar{U}_k} \left\| v^T \frac{1}{n_k n_{\bar{k}}} (x_i - x_j) \right\|_1}{\sum_{k=1}^c \sum_{x_i \in U_k} \sum_{x_j \in U_k} \left\| v^T \frac{1}{n_k n_k} (x_i - x_j) \right\|_1} \\
&= \max_{v^T v=1} \frac{\sum_{k=1}^c \sum_{x_i \in U_k} \sum_{x_j \in \bar{U}_k} \frac{1}{n_k n_{\bar{k}}} |v^T (x_i - x_j)|}{\sum_{k=1}^c \sum_{x_i \in U_k} \sum_{x_j \in U_k} \frac{1}{n_k n_k} |v^T (x_i - x_j)|} \quad (4)
\end{aligned}$$

The objective of the criterion (4) is to find the optimal projection vector v that maximize the ratio of between-class dispersion to the within-class dispersion. These optimized projection vectors are used to construct the optimal projection matrix $V = \{v_1, v_2, \dots, v_d\}$. These projecting vectors are sequentially optimized in d directions. We derive the following iterative algorithm to find the optimal projection vector v that maximizes the objective function (4). The entire algorithmic procedure for L1-SC method is listed below.

B. Algorithmic Procedure for L1-SC

The aforementioned objective function (4) involves maximization of L1-norm based optimization problem. We solve this problem by an iterative algorithm to obtain the optimal projection vector v^* of the matrix V .

The objective function (4) seems similar to the trace ratio formulation of the general graph scaling cut in [3] and [4]. It is difficult to solve (4) by the traditional optimization techniques as both numerator and denominator are constructed by L1-norm maximization and minimization. Inspired by the idea used in [6], [9], [11] and [13], we are using similar L1-norm optimization technique in this work. Thus, we solve the objective function (4) to find the optimal projection vector v^* by the iterative technique. The algorithmic procedure of L1-SC is given as follows.

1. The iteration variable t is set to zero ($t = 0$). Then we randomly initialize the d dimensional vector $v(t)$ and normalize it such that $v(t)^T v(t) = 1$.
2. Two sign functions are defined to compensate the absolute value operation for the numerator and denominator term of (4). These sign functions are computed as

$$q_{ij}(t) = \begin{cases} 1, & \text{if } v^T(t)(x_i - x_j) > 0 \\ -1, & \text{if } v^T(t)(x_i - x_j) \leq 0 \end{cases}$$

and

$$r_{ij}(t) = \begin{cases} 1, & \text{if } v^T(t)(x_i - x_j) > 0 \\ -1, & \text{if } v^T(t)(x_i - x_j) \leq 0 \end{cases} \quad (5)$$

3. Use the sign function to compute $p(t)$ and $b(t)$ by the following equation:

$$\begin{aligned}
p(t) &= \sum_{k=1}^c \sum_{i=1}^{n_k} \sum_{j=1}^{n_{\bar{k}}} q_{ij}(t) \frac{1}{n_k n_{\bar{k}}} v^T(t)(x_i - x_j) \\
b(t) &= \sum_{k=1}^c \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} r_{ij}(t) \frac{1}{n_k n_k} v^T(t)(x_i - x_j) \quad (6)
\end{aligned}$$

Then using $p(t)$ and $b(t)$, update $g(v(t))$

$$g(v(t)) = \frac{p(t)}{v(t)^T p(t)} - \frac{b(t)}{v(t)^T b(t)} \quad (7)$$

4. Then update the vector $v(t)$ using $g(v(t))$ by

$$v(t+1) = v(t) + \gamma g(v(t)) \quad (8)$$

where γ is the learning rate parameter (a small positive value). Then normalize the $v(t+1)$ and update $t = t+1$. If any denominator in (7) happen to zero then perturb the $v(t)$ with a small non zero random vector Δv and update it by $v(t) = (v(t) + \Delta v) / \|(v(t) + \Delta v)\|$ and start with step-2.

5. Convergence check: If the $v(t)$ doesn't show significant increment or $\|v(t+1) - v(t)\| \leq \epsilon$ or total iteration number is greater then maximum given iteration number, then go to step-2 otherwise go to step-6.
6. Stop iteration and assign $v^* = v(t)$.

Above procedure only gives one optimal projection vector. In practical classification problem this one vector is not sufficient for the projection. Hence, It need a projection matrix consists of multiple projection vectors placed in its column space to optimize the objective function. These projection vectors are used to update the input data matrix by

$$X \leftarrow X - v^*(v^{*T})X \quad (9)$$

and then the projection matrix V is padded as $V = [V, v^*]$.

Using the above procedure, we can form the optimal projection matrix V of size $R^{D \times d}$. The pseudo-code for the complete algorithmic procedure for the projection matrix of L1-SC is listed in Algorithm 1.

Algorithm 1: L1-norm based scaling cut algorithm

Input : The training dataset $\{x_i, L_i\}_{i=1}^n \in R^{D \times n}$;
 L_i is the label of each training data x_i ;
Desired dimensionality is d and $d \ll D$.

- 1 Formulate the L1-norm based objective function in (4) to solve the optimization problem.
- 2 Determine the optimal projection vector v^* by solving the optimization problem (4) in Algorithm 2
- 3 Update the input data by using $X = X - v^* v^{*T} X$.
- 4 Pad these optimal projection vectors v^* into the optimal matrix by $V = [V, v^*]$.
- 5 Project the original data into the lower dimensional space d by projection matrix V

Output: Projection matrix $V = \{v_1, v_2, \dots, v_d\} \in R^{D \times d}$, consists of d projection vectors

Result : Projected matrix $Y = V^T X$

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section we evaluate the performance of the proposed L1-SC method on two HSI datasets¹: Salinas ($D = 204, C =$

¹http://www.ehu.es/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes

Algorithm 2: Computation of projection vector for v^*

Input : Number of projection vector d ($d \ll D$);
Learning rate parameter γ ;
Maximum number of iteration is $itmax$

- 1 Set $t = 0$ and Initialize $v(0)$ to a D dimensional random vector such that $v(0)^T v(0) = 1$
 - 2 Compute the sign function $q_{ij}(t)$ and $r_{ij}(t)$ using (5) and set $p(t)$ and $b(t)$ using (6)
 - 3 Determine the $g(v(t))$ function using (7) to update the $v(t)$.
 - 4 Update the $v(t)$ by using (8). where $\gamma > 0$ is the learning parameter.
 - 5 Converge if: $\|v(t+1) - v(t)\| \leq \epsilon$ or $t > itmax$
- Output:** Projection vector $v^* = v(d)$
-

16) and Pavia center ($D = 102, C = 9$). Then we compare it with the state-of-art conventional LDA [2], SC [4], LSC [14] and L1-LDA [6]. In conventional L2-norm based methods use PCA as preprocessing but in L1-norm methods we don't use any preprocessing step. In classification stage, we use SVM classifier with linear kernel to identify the robustness of the proposed algorithm.

In these experiments, we randomly select 10 training samples from each class of the dataset and rest of the samples are used as test dataset. All obtained results are the average of the 5 iterations. Here we evaluate and analyze the effectiveness of the proposed L1-SC by determining the overall classification accuracy of the SVM classifier on the projected data.

Fig. 1 shows the behavior of different L2-norm and L1-norm based algorithms in terms of overall classification accuracy with respect to varied number of input samples. Here, Fig. 1a and Fig. 1b shows the overall classification accuracy of Salinas and Pavia dataset for input data size from 10 to 50. From this figure, it is clearly observed that the proposed L1-SC method completely outperforms the L2-norm based methods and performs on par with L1-LDA when the input data sample size is less.

To better illustrate the noise robustness feature of the proposed L1-SC method with respect to others, we inject white Gaussian noises of different levels to the raw input HSI data and performed the classification on these data using different approaches. The variance of the noise level is varied from 2% to 10% of the variance of the pixel values. As Fig. 2 indicate, the proposed L1-SC method is more robust to noises and achieve better classification accuracies than other methods.

The proposed L1-SC method is compared with other popular L2-norm and L1-norm based methods. The statistics of the highest overall classification accuracy along with corresponding $F1$ -score and dimension of the algorithms for Salinas and Pavia center dataset are highlighted in Table I. Here in Table I, all the results taken classification accuracy are the average of 5 runs using 10 random training samples per class. In order to show the robustness of the algorithm, we have considered the $F1$ -score along with the overall classification accuracy as

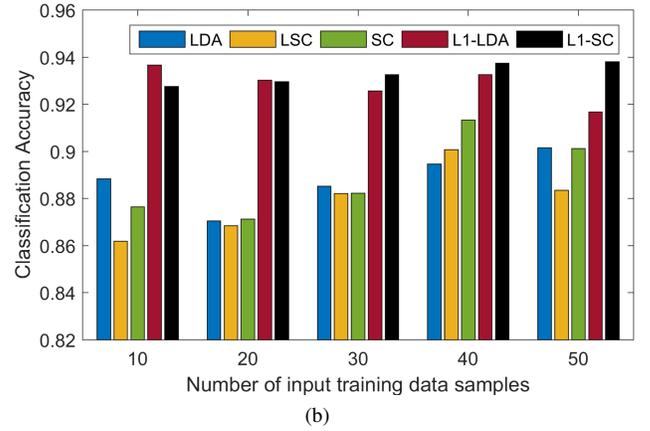
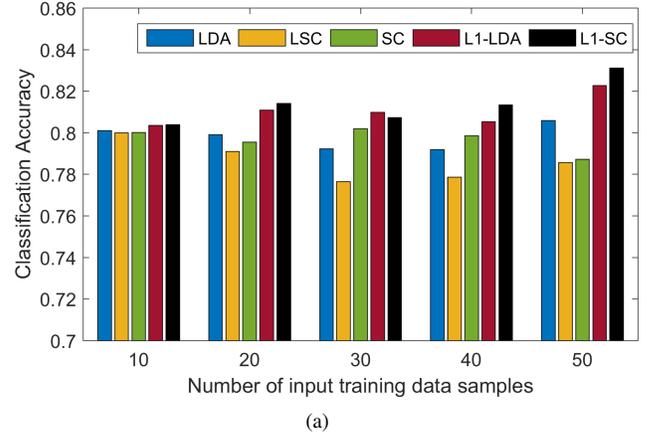


Fig. 1: Effects of different number input samples for the methods on overall accuracies on two data sets Salinas (1a) and Pavia center (1b). From left to right of X-axis shows the overall accuracies with 10, 20, 30, 40 and 50 number input data samples for 10 dimensions.

the performance measure [15]. Table I gives the following observations

- The overall classification accuracy of L1-norm based methods performs better than the other state-of-art L2-norm based methods.
- The classification results of the proposed L1-SC method outperforms the other L2-norm based methods and L1-LDA for both the datasets with less dimensions.
- In Salinas dataset, the proposed L1-SC method produces highest accuracy with maximum $F1$ -score among other approaches by considering only 15 dimensions. Similarly in case of Pavia center dataset, it takes only 10 dimensions. This shows the effectiveness of the algorithm in finding proper projection direction.

The above observations clearly explains the robustness of the proposed algorithm in low dimensional feature space.

TABLE I: Classification performance of proposed approach compared with other L2-norm and L1-norm based approaches

Dataset	Salinas			Pavia Center		
	Methods	Accuracy + stdv	F1-Score	Dims	Accuracy + Stdv	F1-Score
LDA [2]	83.14 ± 1.93	0.8917	35	92.34 ± 0.64	0.8406	25
SC [4]	81.38 ± 1.41	0.8558	40	93.61 ± 0.68	0.8600	25
LSC [14]	83.09 ± 1.88	0.8939	50	93.43 ± 1.07	0.8587	45
L1-LDA [6]	83.21 ± 1.85	0.8913	30	93.50 ± 0.61	0.8542	40
L1-SC (Proposed)	84.01 ± 1.67	0.8956	15	94.20 ± 0.63	0.8724	10

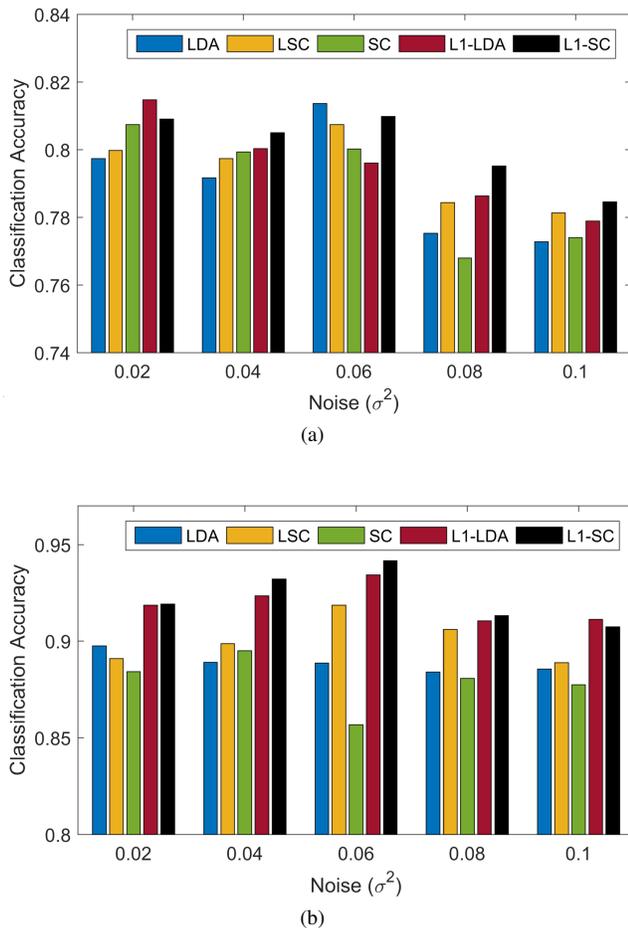


Fig. 2: Illustration of noise robustness of the proposed method with respect to other methods on two data sets Salinas (2a) and Pavia center (2b).

V. CONCLUSION

In this study, we have proposed a novel DR method L1-SC by computing the L1-norm based inter-class and intra-class dispersion. This method determines the projection directions by exploiting the discriminant structure and preserving the geometrical structure of the data. Our method differs from other state-of-art in various ways. For instance, it preserves the intrinsic property as well as the distribution of the data and we believe it handles multimodal and heteroscedastic data with noise and outliers quite well. We examined the performance of our method and other methods over two real world HSI datasets. The promising results of L1-SC on these two datasets

demonstrates its noise robustness and efficiency.

REFERENCES

- [1] A. M. Martínez and A. C. Kak, "Pca versus lda," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228–233, 2001.
- [2] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 8, pp. 982–994, 2004.
- [3] X. Zhang, S. Zhou, and L. Jiao, "Local graph cut criterion for supervised dimensionality reduction," in *Sixth International Symposium on Multispectral Image Processing and Pattern Recognition*. International Society for Optics and Photonics, 2009, pp. 74 9621–74 9621.
- [4] X. Zhang, Y. He, L. Jiao, R. Liu, J. Feng, and S. Zhou, "Scaling cut criterion-based discriminant analysis for supervised dimension reduction," *Knowledge and information systems*, vol. 43, no. 3, pp. 633–655, 2015.
- [5] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 281–288.
- [6] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with l1-norm," *IEEE transactions on cybernetics*, vol. 44, no. 6, pp. 828–842, 2014.
- [7] Y. Liu, Q. Gao, S. Miao, X. Gao, F. Nie, and Y. Li, "A non-greedy algorithm for l1-norm lda," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 684–695, 2017.
- [8] Q. Ke and T. Kanade, "Robust l/sub 1/norm factorization in the presence of outliers and missing data by alternative convex programming," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 739–746.
- [9] N. Kwak, "Principal component analysis based on l1-norm maximization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 9, pp. 1672–1680, 2008.
- [10] X. Li, Y. Pang, and Y. Yuan, "L1-norm-based 2dpc," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 4, pp. 1170–1175, 2010.
- [11] C.-N. Li, Y.-H. Shao, and N.-Y. Deng, "Robust l1-norm two-dimensional linear discriminant analysis," *Neural Networks*, vol. 65, pp. 92–104, 2015.
- [12] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013.
- [13] H. Wang, Q. Tang, and W. Zheng, "L1-norm-based common spatial patterns," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 3, pp. 653–662, 2012.
- [14] X. Zhang, Y. He, N. Zhou, and Y. Zheng, "Semisupervised dimensionality reduction of hyperspectral images via local scaling cut criterion," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 6, pp. 1547–1551, 2013.
- [15] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.