# A GENERALIZED KERNEL RISK SENSITIVE LOSS FOR ROBUST TWO-DIMENSIONAL SINGULAR VALUE DECOMPOSITION

*Miaohua Zhang, Yongsheng Gao, and Jun Zhou*

Institute for Integrated and Intelligent Systems, Griffith University, Australia.

## ABSTRACT

Two-dimensional singular value decomposition (2DSVD) is an important dimensionality reduction algorithm which has inherent advantage in preserving the structure of 2D images. However, 2DSVD algorithm is based on the squared error loss, which may exaggerate the projection errors with the presence of outliers. To solve this problem, we propose a generalized kernel risk sensitive loss for measuring the projection error in 2DSVD, which automatically eliminates the outlier information during optimization. Since the proposed objective function is non-convex, a majorization-minimization algorithm is developed to efficiently solve it. Our method is rotational invariant and has intrinsic advantages in processing non-centered data. Experimental results on public databases demonstrate that the performance of the proposed method significantly outperforms several benchmark methods on different applications.

***Index Terms*—** 2DSVD, kernel, non-convex, robust learning, majorization minimization.

## 1. INTRODUCTION

Principal component analysis (PCA) [1] aims to learn a matrix to project the high-dimensional data to a space with lower dimensionality. However, the traditional PCA algorithm is based on the vector space, which destroys the inherent structure of an image. To better exploit the spatial information carried by image, Yang et al. [2] proposed 2DPCA which directly applies PCA method to 2D images. To improve the robustness of these 2D methods against outliers, Li et al. [3] took the advantages of $L_1$-norm, and proposed $L_1$-2DPCA. Wang et al. [4] proposed a robust 2DPCA based the F-norm to improve its performance when there are outliers in data.

Unlike 2DPCA that employs a one-sided transformation, a two-sided linear transformation was proposed and solved either by an iterative [5] or non-iterative [6] algorithm. Huang and Ding [7] took the rotational invariance property of the $R_1$-norm and applied it to 2DSVD. However, the above methods treat each training sample equally without discriminative constraints for both inliers and outliers [8][9].

Motivated by the success of the information theoretic learning criterions in enhancing the robustness of learning algorithms [10, 11, 12], the kernel risk sensitive loss(KRSL) [13] was proposed for robust adaptive filtering, which was inspired by the risk sensitive loss and maximum correntropy criterion [10]. However, the surface of maximum correntropy criterion is highly non-convex, which means the surface that far away from the optimal solution is flat while the area around the optimal solution is very steep, leading to a suboptimal solution. The KRSL proposed in [13] not only improves the convexity of maximum correntropy criterion but also retains its outlier-resistance ability. However, the KRSL is a kernel based similarity measurement defined in a second order space, which may not always the best choice in matching the representation error. To overcome this problem, we propose a generalized kernel risk sensitive loss (GKRSL) which offers more flexibility in controlling the error, and apply it to the 2DSVD to overcome its vulnerability in dealing with outliers.

The contributions of this paper include (1) a generalized kernel risk sensitive loss (GKRSL) is defined which leads to a robust 2DSVD algorithm for dimensionality reduction; (2) a new majorization minimization optimization procedure is developed to solve the GKRSL-2DSVD; (3) The proposed algorithm is rotational invariant, and the data mean can be automatically updated during the optimization to exclude the information of outliers.

## 2. PROPOSED METHOD

### 2.1. Definition of GKRSL

Given two variables $A$ and $B$, the GKRSL is given by

$$\begin{aligned} f_{\text{GKRSL}}(A-B) &= \frac{1}{\lambda}\mathbf{E}\left[\exp\left(\lambda\eta\|\kappa(A)-\kappa(B)\|_H^p\right)\right] \\ &= \frac{1}{\lambda}\mathbf{E}\left[\exp\left(\lambda\eta\left(\|\kappa(A)-\kappa(B)\|_H^2\right)^{\frac{p}{2}}\right)\right] \\ &= \frac{1}{\lambda}\mathbf{E}[\exp(\lambda(1-g_\sigma(A-B))^{\frac{p}{2}})], \end{aligned} \tag{1}$$

where $\eta = 2^{-\frac{p}{2}}$, $p > 0$ is the order of error loss [13, 14], and GKRSL reduces to KRSL when $p$ is 2. $\lambda > 0$ is a parameter that controls the convexity of the function, $\mathbf{E}(x)$ is the expectation of $x$. $g_\sigma(x)$ is a Mercer kernel with the bandwidth being $\sigma$. $\kappa(x)$ is a nonlinear mapping function that maps the variable $x$ from the original space to the kernel spaces, thus Eq.(1) can also be regarded as a similarity measurement between $A$

and $B$ in the kernel space $H$. In actual implementation, only a finite number of samples are available, the Parzen windowing method can be applied to estimate the GKRSL on a finite number of available samples $\{(a_i, b_i)\}_{i=1}^N$ [15, 16]:

$$f_{\text{GKRSL}}(A-B) = \frac{1}{N\lambda}\sum_{i=1}^{N}\exp(\lambda(1-k_\sigma(a_i-b_i))^{\frac{p}{2}}). \quad (2)$$

### 2.2. The Proposed GKRSL-2DSVD

Compared with MCC and KRSL, the GKRSL gives a more flexible choice in controlling the representation error, and thus its error fitting ability will be much enhanced. Therefore, we propose the following GKRSL model to learn more robust features for 2DSVD with the presence of outliers.

$$\begin{aligned}&\min_{L,R,\{M_i\},\bar{X}} f_{\text{GKRSL}}(E_i) = \frac{1}{N\lambda}\sum_{i=1}^{N}\exp(\lambda(1-\exp(-\frac{E_i^2}{2\sigma^2}))^{\frac{p}{2}})\\ &\text{s.t. } L^TL=I,\ \ R^TR=I,\ \ E_i=\sqrt{\|\hat{X}_i - LM_iR^T\|_F^2}.\end{aligned} \quad (3)$$

We first calculate $M_i$ by setting the derivative of $f_{\text{GKRSL}}$ with respect to $M_i$ to zero:

$$\begin{aligned}\frac{\partial f_{\text{GKRSL}}}{\partial M_i} &= \frac{p\lambda}{2\sigma_2}Q_1Q_2Q_3(X_i - LM_iR^T)L^TR = 0.\\ Q_1 &= \exp(\lambda(1-\exp(-\frac{E_i^2}{2\sigma^2}))^{\frac{p}{2}}),\\ Q_2 &= (1-\exp(-\frac{E_i^2}{2\sigma^2}))^{\frac{p}{2}-1},\ \ Q_3 = \exp(-\frac{E_i^2}{2\sigma^2}).\end{aligned} \quad (4)$$

Since $Q_1$,$Q_2$, and $Q_3$ are all positive, the term $X_i - LM_iR^T$ should be zero, then we have $M_i = L^TX_iR$. Thus Eq.(3) can be rewritten as

$$\begin{aligned}&\min_{L,R,\bar{X}} f_{\text{GKRSL}}(E_i) = \frac{1}{N\lambda}\sum_{i=1}^{N}\exp(\lambda(1-\exp(-\frac{E_i^2}{2\sigma^2}))^{\frac{p}{2}})\\ &\text{s.t. } L^TL=I, R^TR=I, E_i=\sqrt{\|\hat{X}_i - LL^T\hat{X}_iRR^T\|_F^2}.\end{aligned} \quad (5)$$

Since Eq.(5) is non-convex, we introduce a majorization minimization algorithm to solve this nonconvex optimization problem, which consists of the following two steps:
(1) construct a convex upper bound surrogate function for the non-convex objective function, i.e., $f_{\text{GKRSL}}(E|E_t)$.
(2) minimize the function $f_{\text{GKRSL}}(E|E_t)$ until convergence.

### 2.3. Majorization Procedure

Here we introduce how to construct the surrogate function. Since the function $f_{\text{GKRSL}}(E)$ is non-decreasing and non-convex, the first Taylor expansion of $f_{\text{GKRSL}}(E)$ in the proximity point $E_{i,t}$ satisfies

$$\begin{aligned}f_{\text{GKRSL}}(E_i) &\le f_{\text{GKRSL}}(E_{i,t}) + f'(E_{i,t})(E_i - E_{i,t}) + c\\ &= f_{\text{GKRSL}}(E_i|E_{i,t}),\end{aligned} \quad (6)$$

where $c$ is a constant, and $t$ is the iteration number [17].

According to the MM theory in [18], we have

$$f(E) \le f_{\text{GKRSL}}(E|E_t), \ \text{ and } \ f_{\text{GKRSL}}(E_t) = f_{\text{GKRSL}}(E_t|E_t). \quad (7)$$

If the $E_{t+1}$ denotes the minimizer of the $f_{\text{GKRSL}}(E|E_t)$, then the MM procedure has the descent property as

$$f_{\text{GKRSL}}(E_{t+1}) \le f_{\text{GKRSL}}(E_t), \qquad t = 1, 2, \cdots. \quad (8)$$

Then the objective function can be upperbounded by $f'_{\text{GKRSL}}(E_t)E$ by omitting the constant terms in $f_{\text{GKRSL}}(E|E_t)$

$$\min f_{\text{GKRSL}}(E) \le f'_{\text{GKRSL}}(E_t)E. \quad (9)$$

### 2.4. Minimization Procedure

Based on the above analysis, minimizing Eq.(5) can be achieved by minimizing the following surrogate function

$$\begin{aligned}&\operatorname*{argmin}_{L,R,\bar{X}} \ f_{\text{GKRSL}}(E|E_t)\\ &\text{s.t. } L^TL=I,\ \ R^TR=I, E_i=\sqrt{\|\hat{X}_i - LL^T\hat{X}_iRR^T\|_F^2}.\end{aligned} \quad (10)$$

The Lagrangian function of Eq.(10) is

$$\begin{aligned}\mathcal{L}(\hat{L},\hat{R},\hat{\bar{X}}) &= f_{\text{GKRSL}}(E|E_t)\\ &+ \text{Tr}(\Omega_1(L^TL-I)) + \text{Tr}(\Omega_2(R^TR-I)),\end{aligned} \quad (11)$$

where $\text{Tr}(x)$ is the trace of $x$. According to Eq.(6), we have

$$f_{\text{GKRSL}}(E|E_t) = f'_{\text{GKRSL}}(E_t)E = \frac{p}{2}Q_1Q_2Q_3E_tE \quad (12)$$

Let $W = \frac{p}{2}Q_1Q_2Q_3E_t$ be the weight for each sample. Thus Eq.(11) can be rewritten as

$$\begin{aligned}&\operatorname*{argmin}_{\hat{L},\hat{R},\hat{\bar{X}}} \ \mathcal{L}\{\hat{L},\hat{R},\hat{\bar{X}}\}\\ &= \frac{1}{N}\sum_{i=1}^{N}W_i\sqrt{\text{Tr}(\hat{X}_i^T\hat{X}_i - \hat{X}_i^TLL^T\hat{X}_iRR^T)}\\ &\quad + \text{Tr}(\Omega_1(L^TL-I)) + \text{Tr}(\Omega_2(R^TR-I)),\\ &\text{s.t. } \ W_i = \frac{p}{2}Q_1Q_2Q_3E_{i,t}, \quad \hat{X}_i = X - \bar{X}.\end{aligned} \quad (13)$$

The optimal solution $\{\hat{L},\hat{R},\hat{\bar{X}}\}$ can be obtained by setting the derivative of Lagrangian function in Eq.(13) with respect to (w.r.t.) $L$, $R$, and $\hat{\bar{X}}$, respectively. First, the optimal $\hat{\bar{X}}$ can be obtained by solving the following problem:

$$\frac{\partial\mathcal{L}}{\partial\bar{X}} = \frac{\partial\sum_{i=1}^{N}W_i\sqrt{\|\hat{X}_i - LL^T\hat{X}_iRR^T\|_F^2}}{\partial\bar{X}} = 0 \quad (14)$$

By solving Eq.(14), the optimal $\hat{\bar{X}}$ can obtained by

$$\hat{\bar{X}} = \sum_{i=1}^{N}\frac{\frac{1}{2}W_iX_i}{\sqrt{\text{Tr}(\hat{X}_i^T\hat{X}_i - \hat{X}_i^TLL^T\hat{X}_iRR^T)}}/\sum_{i=1}^{N}W_i. \quad (15)$$

We solve the optimal $\hat{L}$ and $\hat{R}$ by setting the derivative of Lagrangian function w.r.t. $L$ and $R$ as follows.

$$\frac{\partial \mathcal{L}}{\partial L} = -FL + \Omega_1 L = 0, \quad \frac{\partial \mathcal{L}}{\partial R} = -GR + \Omega_2 R = 0, \tag{16}$$

where $F$ is calculated by $F = \frac{W_i \hat{X}_i R R^T \hat{X}_i^T}{2\sqrt{\mathrm{Tr}(\hat{X}_i^T \hat{X}_i - \hat{X}_i^T L L^T \hat{X}_i R R^T)}}$, the optimal $\hat{L}$ can be updated by the largest $k_1$ eigenvectors of $F$. $G$ is calculated by $G = \frac{W_i \hat{X}_i^T L L^T \hat{X}_i}{2\sqrt{\mathrm{Tr}(\hat{X}_i^T \hat{X}_i - \hat{X}_i^T L L^T \hat{X}_i R R^T)}}$, and the optimal $\hat{R}$ can be updated by the largest $k_2$ eigenvectors of $G$. This alternate optimization procedure is repeated until the error between the current $L$ and $R$ and the ones calculated in the last iteration falls below a prescribed threshold $\epsilon$.

As with other kernel methods, kernel size (bandwidth) selection will affect the performance of the proposed method, whose value is often determined empirically [14][16]. In this work, the kernel size is determined by

$$\sigma^2 = \frac{1}{2N} \sum_{i=1}^{N} \mathrm{Tr}(\hat{X}_i^T \hat{X}_i - \hat{X}_i^T \hat{L} \hat{L}^T \hat{X}_i \hat{R} \hat{R}^T). \tag{17}$$

Based on the above analysis, the robust GKRSL-2DSVD algorithm is summarized in Algorithm 1.

**Algorithm 1** GKRSL-2DSVD Algorithm

**Input:** Given a data matrix $X = \{X_1, X_2, \cdots, X_N\}$ with each $X_i \in R^{m \times n}$. $p$, $\lambda$, $k_1$, $k_2$, and threshold $\epsilon$.

**Output:** $\{W_i\}_{i=1}^N$, $\hat{L}, \hat{R}, \hat{\hat{X}}$.

**while** $t$ = 1,..., $t$ **do**
  Update weight $\{W_i\}_{i=1}^N$ for each sample by Eq.(13).
  Update $\hat{\hat{X}}$ by Eq.(15).
  Update $\hat{L}$ and $\hat{R}$ by Eq.(16).
  1) Update $F$ using the current $L$ and $R$, the optimal $\hat{L}$ is the largest $k_1$ eigenvectors of $F$,
  2) Update $G$ using the current $L$ and $R$, the optimal $\hat{R}$ is the largest $k_2$ eigenvectors of $G$,
  Update $\sigma^2$ by Eq.(17)
  **if** $\epsilon > 1e-5$ **then**
    repeat;
  **else**
    $t \leftarrow t+1$; Break;
  **end if**
**end while**

## 3. EXPERIMENTAL RESULTS

### 3.1. Datasets and Parameter Settings

To verify the effectiveness of the proposed algorithm, we carry out experiments on MNIST[1] dataset and ORL Face Dataset[2] [19, 20] for image classification and clustering. The proposed algorithm is tested via different evaluation measurements and compared with seven classical 2D subspace learing algorithms, including 2DPCA [2], $L_1$-2DPCA [3], F-2DPCA [4], 2DSVD [6], $R_1$-2DSVD [7], N-2DNPP [21], and S-2DNPP [21]. In the proposed algorithm, $\lambda$ controls the convex range and $p$ controls the error distribution. In all our experiments, $\lambda$ and $p$ are both empirically set to 8. All the experiments are conducted on MATLAB R2015a.

**Table 1**. The recognition accuracy of all the algorithms on the MNIST dataset with 5% outliers: Average recognition accuracy (AC) $\pm$ standard derivation.

| Methods | Images per digit $\times$ ♯ of digits | | |
|---|---|---|---|
| | $400 \times 10$ | $600 \times 10$ | $800 \times 10$ |
| 2DPCA | $0.6264 \pm 0.0206$ | $0.6543 \pm 0.0171$ | $0.6788 \pm 0.0136$ |
| $L_1$-2DPCA | $0.6257 \pm 0.0204$ | $0.6539 \pm 0.0171$ | $0.6782 \pm 0.0136$ |
| F-2DPCA | $0.6272 \pm 0.0164$ | $0.6490 \pm 0.0119$ | $0.6759 \pm 0.0122$ |
| 2DSVD | $0.6360 \pm 0.0160$ | $0.6565 \pm 0.0121$ | $0.6840 \pm 0.0113$ |
| $R_1$-2DSVD | $0.6358 \pm 0.0162$ | $0.6562 \pm 0.0121$ | $0.6562 \pm 0.0121$ |
| N-2DNPP | $0.6405 \pm 0.0130$ | $0.6548 \pm 0.0160$ | $0.6689 \pm 0.0131$ |
| S-2DNPP | $0.6283 \pm 0.0213$ | $0.6566 \pm 0.0154$ | $0.6799 \pm 0.0136$ |
| Proposed | $\mathbf{0.8462 \pm 0.0041}$ | $\mathbf{0.8458 \pm 0.0014}$ | $\mathbf{0.8639 \pm 0.0020}$ |

### 3.2. Experiments on Image Classification

Here we test our algorithms on the MNIST dataset with the presence of outliers. We respectively choose 400, 600, 800 samples per digit from the training set for training, and use all the testing samples for testing. All the samples are normalized by their norms. To simulate outliers, we randomly choose 5% of the training samples and weight them by a magnitude $a$. i.e., $X_o = aX_c$ where $X_o$ and $X_c$ denote the simulated outlier image and clean image. We first set the magnitude ($a$) of the outliers to 50 and number of principal components to $k_1 = k_2 = 15$ to evaluate the proposed method under varying number of training samples. 1 nearest neighbor (1NN) is used as the classifier for all the algorithms. The classification accuracies of different algorithms using the above settings are listed in Table 1. All the results are reported over 20 random trials to reduce the statistical deviation. The results show that the proposed algorithm outperforms all the benchmarks on different sizes of training samples.

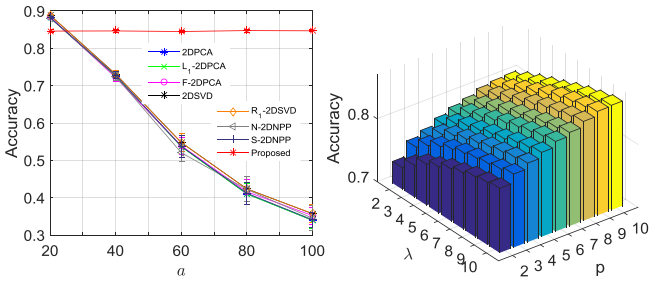
To check the influence of varying magnitude of outliers on the accuracy, we test all the algorithms on the $400 \times 10$ training samples, and vary $a$ from 20 to 100. The accuracies are shown in Fig. 1 (a), from which we can see that the performance of the proposed algorithm are almost unaffected under different $a$ while the accuracies of other algorithms reduces rapidly when $a$ increases. We also plot a barchart in Fig. 1(b) showing the classification accuracy under different $\lambda$ and $p$. We can see that, with a fixed $p$ value, the accuracy increases with the $\lambda$ increasing. When the $\lambda$ is set to a fixed value, the accuracies increase fast with the increase of $p$.

[1] http://yann.lecun.com/exdb/mnist/

[2] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

**Table 2**. Clustering results of subspaces learned from different algorithms on the first 100 faces of the ORL dataset: Average Clustering Accuracy (AC) $\pm$ Standard Deviation and Average normalized mutual information (NMI) $\pm$ Standard Deviation.

| Methods and evaluation metrics | | Number of principal components | | | |
|---|---|---|---|---|---|
| | | $m = 30$ | $m = 50$ | $m = 70$ | $m = 90$ |
| 2DPCA | AC | 0.5991 $\pm$ 0.0442 | 0.7535 $\pm$ 0.0153 | 0.8143 $\pm$ 0.0190 | 0.7507 $\pm$ 0.0070 |
| | NMI | 0.7619 $\pm$ 0.0268 | 0.8692 $\pm$ 0.0042 | 0.8860 $\pm$ 0.0052 | 0.8684 $\pm$ 0.0019 |
| $L_1$-2DPCA | AC | 0.6981 $\pm$ 0.0176 | 0.8199 $\pm$ 1.2e-15 | 0.8003 $\pm$ 0.0315 | 0.7500 $\pm$ 0 |
| | NMI | 0.8221 $\pm$ 0.0112 | 0.8875 $\pm$ 1.4e-15 | 0.8821 $\pm$ 0.0087 | 0.8682 $\pm$ 4.4e-16 |
| F-2DPCA | AC | 0.7000 $\pm$ 1.3e-15 | 0.8199 $\pm$ 1.2e-15 | 0.7528 $\pm$ 0.0137 | 0.7500 $\pm$ 0 |
| | NMI | 0.8200 $\pm$ 7.8e-16 | 0.8875 $\pm$ 1.4e-15 | 0.8690 $\pm$ 0.0038 | 0.8682 $\pm$ 4.4e-16 |
| 2DSVD | AC | 0.7012 $\pm$ 0.0836 | 0.7571 $\pm$ 0.0219 | 0.8108 $\pm$ 0.0236 | 0.7528 $\pm$ 0.0137 |
| | NMI | 0.8197 $\pm$ 0.0417 | 0.8615 $\pm$ 0.0136 | 0.8850 $\pm$ 0.0065 | 0.8690 $\pm$ 0.0038 |
| $R_1$-2DSVD | AC | 0.6876 $\pm$ 0.0781 | 0.7615 $\pm$ 0.0165 | 0.8052 $\pm$ 0.0286 | 0.7507 $\pm$ 0.0070 |
| | NMI | 0.8128 $\pm$ 0.0406 | 0.8640 $\pm$ 0.0095 | 0.8835 $\pm$ 0.0079 | 0.8684 $\pm$ 0.0019 |
| N-2DNPP | AC | 0.7975 $\pm$ 0.0925 | 0.7822 $\pm$ 0.0351 | 0.7948 $\pm$ 0.0338 | 0.7528 $\pm$ 0.0138 |
| | NMI | 0.8753 $\pm$ 0.0295 | 0.8772 $\pm$ 0.0097 | 0.8806 $\pm$ 0.0093 | 0.8691 $\pm$ 0.0038 |
| S-2DNPP | AC | 0.7411 $\pm$ 0.0250 | 0.7424 $\pm$ 0.0129 | 0.8163 $\pm$ 0.0177 | 0.7491 $\pm$ 0.0090 |
| | NMI | 0.8223 $\pm$ 0.0148 | 0.8457 $\pm$ 0.0070 | 0.8859 $\pm$ 0.0082 | 0.8666 $\pm$ 0.0065 |
| Proposed | AC | **0.9160 $\pm$ 0.0479** | **0.9377 $\pm$ 0.0363** | **0.8461 $\pm$ 0.0721** | **0.7623 $\pm$ 0.0258** |
| | NMI | **0.9158 $\pm$ 0.0241** | **0.9292 $\pm$ 0.0191** | **0.8902 $\pm$ 0.0332** | **0.8704 $\pm$ 0.0078** |



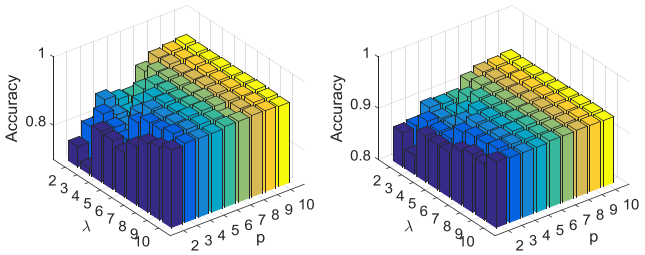

(a) (b)

**Fig. 1**. AC on the MNIST Handwritten Digit Dataset. (a) AC of all the algorithms with changing magnitude of outliers; (b) AC of the proposed algorithm with different $\lambda$ and $p$.

### 3.3. Experiments on Image Clustering

We then test the proposed algorithm and all the benchmarks on an image clustering problem on the ORL face dataset with the presence of outliers. We select all the face images of the first 10 subjects as the training samples, i.e., 100 images are selected as the training data. We randomly generate 30 dummy images as outliers [7][22] and add them to the training data, thus the number of clean training and outlier samples are 100 and 30, respectively. After learning the features by all the algorithms, K-means algorithm is applied to evaluate the quality of these features. Before applying the K-means, we initialize the clustering center by the density search based method proposed in [23].

To apply K-means, for the one-sided transforms, including 2DPCAs and 2DLPPs, we directly apply K-means on the projected samples, i.e., $X_i^{\text{new}} = \hat{X}_i U$, where $U$ is the projection matrix and $i = 1, 2, \cdots, N$. For the two-sided transforms, we apply the K-means to the $\{M_i\}_{i=1}^N$. The clustering performance is measured by average clustering accuracy (AC) and average normalized mutual information (NMI) [24]. The results of different algorithm under varying number of principal components ($m$) are shown in Table 2 where all the results are reported over 100 iterations. These results show that the performance of the proposed algorithm is the best under different $m$. We then plot AC and NMI with varying $\lambda$ and $p$ in Fig. 2. We can see that, better AC and NMI can be obtained by choosing a $p$ larger than 2. When $p$ is fixed, the clustering AC and NMI increase with $\lambda$ increasing.




(a) (b)

**Fig. 2**. AC with different $\lambda$ and $p$ on the ORL dataset. (a) AC; (b) NMI.

## 4. CONCLUSION

In this paper, we developed a generalized kernel risk sensitive loss for robust 2DSVD (GKRSL-2DSVD) which can discriminatively weight the training samples so that the information of the outliers is excluded from the training procedure. Thus the learned features from the proposed model are more robust to outliers. Since the resulted objective function is a non-convex, we developed an optimization algorithm based on the majorization minimization theory. Extensive experiments on two image processing tasks on public datasets with varying parameter settings show that the proposed algorithm has superior outlier-resistance ability to other benchmarks.

## 5. REFERENCES


[1] Matthew Turk and Alex Pentland, "Eigenfaces for recognition," *J. cogn. neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[2] Jian Yang, David Zhang, Alejandro F Frangi, and Jingyu Yang, "Two-dimensional PCA: a new approach to appearance-based face representation and recognition," *IEEE Trans Pattern Anal Mach Intell.*, vol. 26, no. 1, pp. 131–137, 2004.

[3] Xuelong Li, Yanwei Pang, and Yuan Yuan, "$L_1$-norm-based 2DPCA," *IEEE Trans. Syst. Man Cybern. B Cybern.*, vol. 40, no. 4, pp. 1170–1175, 2010.

[4] Qianqian Wang and Quanxue Gao, "Two-dimensional PCA with F-norm minimization," in *AAAI*, 2017.

[5] Jieping Ye, "Generalized low rank approximations of matrices," *Mach. Learn.*, vol. 61, no. 1-3, pp. 167–191, 2005.

[6] Chris Ding and Jieping Ye, "2-Dimensional singular value decomposition for 2D maps and images," in *SIAM Int'l Conf. Data Mining,*, 2005, pp. 32–43.

[7] Heng Huang and Chris Ding, "Robust tensor factorization using $R_1$ norm," in *Comput. Vis. Pattern Recognit. (CVPR)*, 2008, pp. 1–8.

[8] Miaohua Zhang, Yongsheng Gao, Changming Sun, and Michael Blumenstein, "Kernel mean $p$ power error loss for robust two-dimensional singular value decomposition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2019, pp. 3432–3436.

[9] Miaohua Zhang, Yongsheng Gao, Changming Sun, John La Salle, and Junli Liang, "Robust tensor factorization using maximum correntropy criterion," in *Proc. IEEE Int. Conf. Pattern Recognit. (ICPR)*. IEEE, 2016, pp. 4184–4189.

[10] Ran He, Bao-Gang Hu, Wei-Shi Zheng, and Xiang-Wei Kong, "Robust principal component analysis based on maximum correntropy criterion," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1485–1494, 2011.

[11] Miaohua Zhang, Yongsheng Gao, Changming Sun, and Michael Blumenstein, "A robust matching pursuit algorithm using information theoretic learning," *Pattern Recognit.*, vol. 107, pp. 107415, 2020.

[12] Miaohua Zhang, Yongsheng Gao, Changming Sun, and Michael Blumenstein, "Robust sparse learning based on kernel non-second order minimization," in *IEEE Int. Conf. Image Process. (ICIP)*. IEEE, 2019, pp. 2045–2049.

[13] Badong Chen, Lei Xing, Bin Xu, Haiquan Zhao, Nanning Zheng, and Jose C Principe, "Kernel risk-sensitive loss: definition, properties and application to robust adaptive filtering," *IEEE Trans. Signal Process.*, vol. 65, no. 11, pp. 2888–2901, 2017.

[14] Lei Xing, Yunqi Mi, Yuanhao Li, and Badong Chen, "Robust locality preserving projection based on kernel risk-sensitive loss," in *IJCNN*. IEEE, 2018, pp. 1–7.

[15] Weifeng Liu, Puskal P Pokharel, and José C Príncipe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, 2007.

[16] Ran He, Wei-Shi Zheng, and Bao-Gang Hu, "Maximum correntropy criterion for robust face recognition," *IEEE Trans Pattern Anal Mach Intell.*, vol. 33, no. 8, pp. 1561–1576, 2011.

[17] Miaohua Zhang, Yongsheng Gao, and Jun Zhou, "A unified weight learning and low-rank regression model for robust complex error modeling," *Pattern Recognit.*, vol. 120, pp. 108147, 2021.

[18] David R Hunter and Kenneth Lange, "A tutorial on MM algorithms," *Am. Stat.*, vol. 58, no. 1, pp. 30–37, 2004.

[19] Xiaozheng Zhang and Yongsheng Gao, "Face recognition across pose: A review," *Pattern recognit.*, vol. 42, no. 11, pp. 2876–2896, 2009.

[20] Yongsheng Gao and Yutao Qi, "Robust visual similarity retrieval in single model face databases," *Pattern Recognit.*, vol. 38, no. 7, pp. 1009–1020, 2005.

[21] Zhao Zhang, Fanzhang Li, Mingbo Zhao, Li Zhang, and Shuicheng Yan, "Robust neighborhood preserving projection by nuclear/$L_{2,1}$-norm regularization for image feature extraction," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1607–1622, 2017.

[22] Miaohua Zhang, Yongsheng Gao, Changming Sun, and Michael Blumenstein, "Robust tensor decomposition for image representation based on generalized correntropy," *IEEE Trans. Image Process.*, vol. 30, pp. 150–162, 2020.

[23] Alex Rodriguez and Alessandro Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[24] Deng Cai, Xiaofei He, and Jiawei Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data En.*, vol. 17, no. 12, pp. 1624–1637, 2005.