Griffith UNIVERSITY
Queensland, Australia

22-27 MAY 2022
ICASSP 2022
SINGAPORE

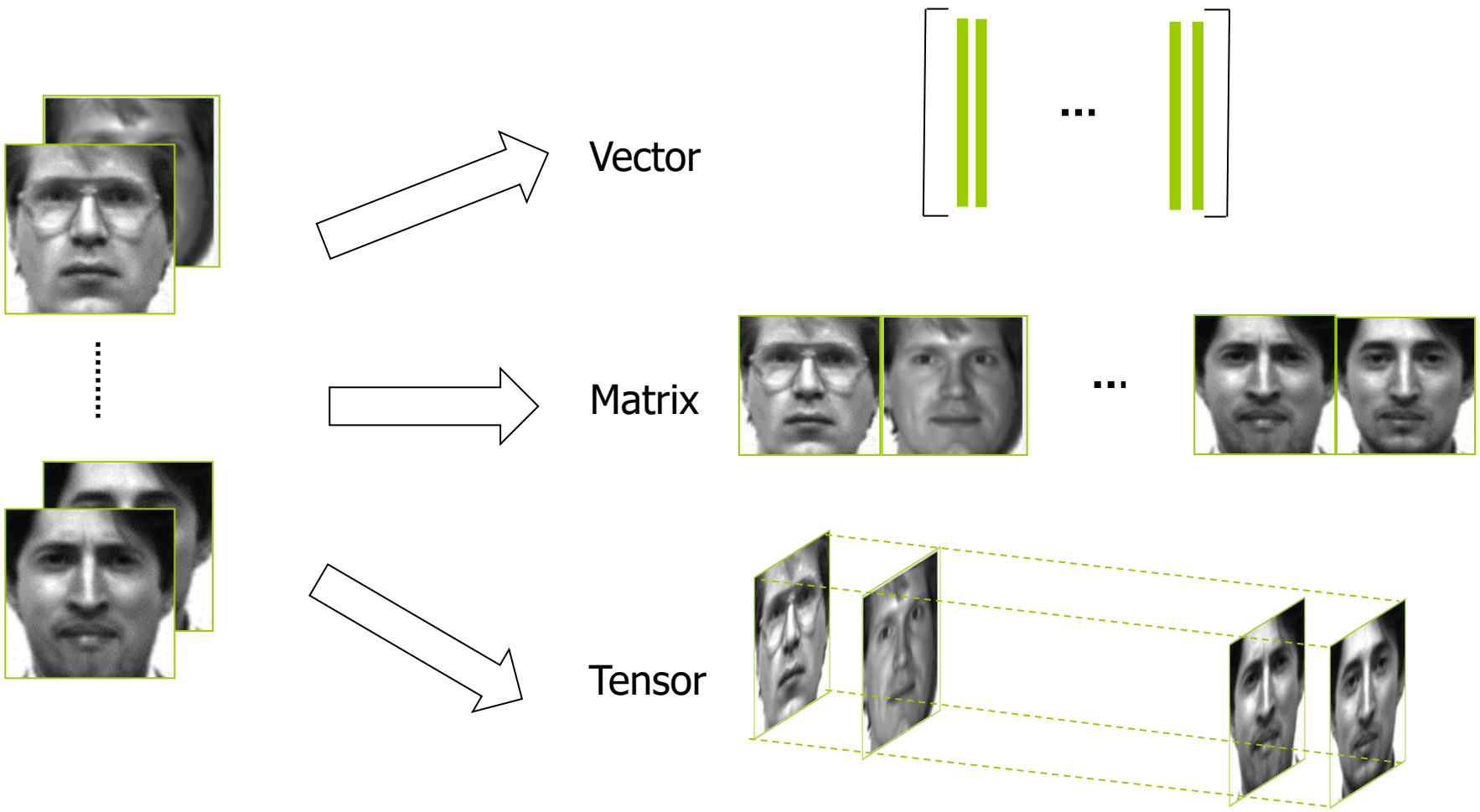
A Generalized Kernel Risk Sensitive Loss for Robust Two-dimensional Singular Value Decomposition

Miaohua Zhang, Yongsheng Gao, Jun Zhou

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

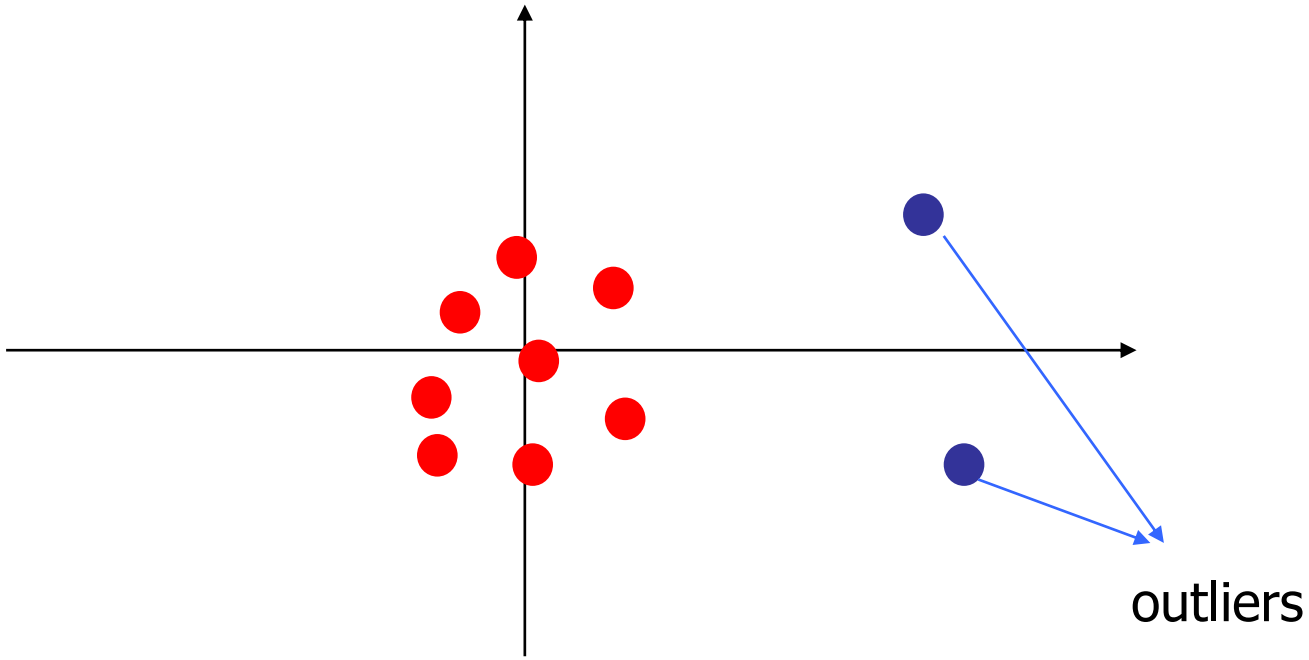
1. Introduction and Motivation

- Representation of Data



1. Introduction and Motivation

- Data with outliers



2. Background

- Minimum error formulation of PCA

Matrix: $\left\|\mathbf{X}-\mathbf{UV}\right\|_F^2$

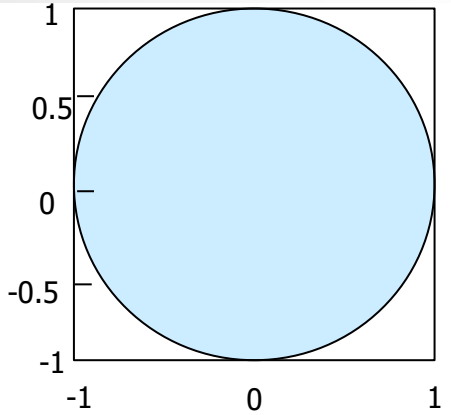
Vector: $\min_{\mathbf{U},\mathbf{V},\boldsymbol{\mu}} \sum_i \left\|\mathbf{x}_i-\left(\boldsymbol{\mu}+\mathbf{U}\mathbf{v}_i\right)\right\|_2^2$

Mean square error(MSE)

2. Background

- L_1 norm based PCA

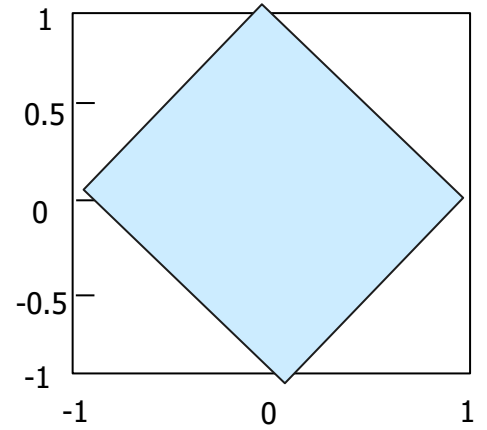
$$\|\mathbf{X}\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{d} \mathbf{x}_{ji}^2 \right)^{1/2}, \quad \|\mathbf{X}\|_1 = \sum_{i=1}^{n} \sum_{j=1}^{d} |\mathbf{x}_{ji}|$$



L_2 norm based curve

L_1 Norm PCA has the cost function

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{X} - \mathbf{UV}\|_{L_1} = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{U}\mathbf{v}_i\|_{L_1}$$



L_1 norm based curve

2. Background

- Two-dimensional Principle component analysis

Two dimensional approaches directly apply decomposition on 2D images. For example, 2DPCA uses all the 2D images to construct a covariance matrix : $\mathbf{C}$

$$\mathbf{C} = \sum_{i=1}^{N} (\mathbf{X}_i - \boldsymbol{\mu})^T (\mathbf{X}_i - \boldsymbol{\mu}) = \sum_{i=1}^{N} \mathbf{X}_i^{\boldsymbol{\mu}T} \mathbf{X}_i^{\boldsymbol{\mu}}$$

Where $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i$

Jian Yang, David Zhang, Alejandro F Frangi, and Jing-yu Yang. Two-dimensional pca: a new approach to appearance-based face representation and recognition. IEEE transactions on pattern analysis and machine intelligence, 26(1):131–137, 2004.

2. Background

- Two-dimensional singular vector decomposition

Objective function:

$$\min_{L,R,\{M_i\}} J(\mathbf{L},\{\mathbf{M}_i\},\mathbf{R}) = \sum_{i=1}^{N} \left\| \mathbf{X}_i^{\mu} - \mathbf{L}\mathbf{M}_i\mathbf{R}^T \right\|_F^2$$

Where $\mathbf{L} \in \Re^{p\times k_1}, \mathbf{R} \in \Re^{q\times k_2}$ and $\mathbf{M}_i \in \Re^{k_1\times k_2}$. Define the row-row and column-column covariance matrices as:

$$\mathbf{C}_1 = \sum_{i=1}^{N} \mathbf{X}_i^{\mu}\mathbf{R}\mathbf{R}^T\mathbf{X}_i^{\mu T} \qquad \mathbf{C}_2 = \sum_{i=1}^{N} \mathbf{X}_i^{\mu T}\mathbf{L}\mathbf{L}^T\mathbf{X}_i^{\mu}$$

The 2DPCA is a special case of 2DSVD by setting $\mathbf{L} = \mathbf{I}$

3. Proposed Method

We developed a generalized kernel risk sensitive loss for robust 2DSVD decomposition.

Definition: Generalized kernel risk sensitive loss (GKRSL)

- Based on the information potential, the GKRISL is defined as a generalized similarity measure between two arbitrary random variables A and B

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

3. Proposed Method

Advantages:

- It is a local criterion, compared with the global criterion, like mean square error, the local criterion will be more accurate.
- Compared with the second order metric, the P-order metric offers more flexible choice in controlling the representation error.
- The optimization of generalized kernel risk sensitive loss is easier than that of L1-norm based methods.
- It has a clear theoretical foundation and it satisfies symmetric, positive, triangle inequality and rotational invariant.

3. Proposed Method

- The robust 2DSVD based on GKRSL

$$\min_{L,R,\{M_i\},\bar{X}} f_{\mathrm{GKRSL}}(E_i) = \frac{1}{N\lambda}\sum_{i=1}^{N}\exp(\lambda(1-\exp(-\frac{E_i^2}{2\sigma^2}))^{\frac{p}{2}})$$

$$\text{s.t. } L^TL = I,\ \ R^TR = I,\ \ E_i = \sqrt{\|\hat{X}_i - LM_iR^T\|_F^2}.$$

First, we solve the optimization on $\mathbf{M}_i$ by setting the derivative of loss function with respect to $\mathbf{M}_i$ to zeros

$$\min_{L,R,\bar{X}} f_{\mathrm{GKRSL}}(E_i) = \frac{1}{N\lambda}\sum_{i=1}^{N}\exp(\lambda(1-\exp(-\frac{E_i^2}{2\sigma^2}))^{\frac{p}{2}})$$

$$\text{s.t. } L^TL = I, R^TR = I, E_i = \sqrt{\|\hat{X}_i - LL^T\hat{X}_iRR^T\|_F^2}.$$

3. Proposed Method

- Optimization by Majorization Minimization

Majorization step: construct a convex upper bound surrogation function for the non-convex objective function.

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

$$\min f_{\mathrm{GKRSL}}(E) \leq f'_{\mathrm{GKRSL}}(E_t)E.$$

3. Proposed Method

- Optimization by Majorization Minimization

Minimization step: Minimize the surrogate function until convergence

$$\underset{L,R,\bar{X}}{\operatorname{argmin}} \quad f_{\mathrm{GKRSL}}(E|E_t)$$

$$\text{s.t. } L^T L = I, \quad R^T R = I, E_i = \sqrt{\|\hat{X}_i - LL^T \hat{X}_i RR^T\|_F^2}.$$

$$\mathcal{L}(\hat{L}, \hat{R}, \hat{\bar{X}}) = f_{\mathrm{GKRSL}}(E|E_t) + \mathrm{Tr}(\Omega_1(L^T L - I)) + \mathrm{Tr}(\Omega_2(R^T R - I)),$$

3. Proposed Method

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

$$\hat{\bar{X}} = \sum_{i=1}^{N} \frac{\frac{1}{2} W_i X_i}{\sqrt{\mathrm{Tr}(\hat{X}_i^T \hat{X}_i - \hat{X}_i^T LL^T \hat{X}_i RR^T)}} / \sum_{i=1}^{N} W_i.$$

$$\frac{\partial \mathcal{L}}{\partial L} = -FL + \Omega_1 L = 0,$$

$$F = \frac{W_i \hat{X}_i RR^T \hat{X}_i^T}{2\sqrt{\mathrm{Tr}(\hat{X}_i^T \hat{X}_i - \hat{X}_i^T LL^T \hat{X}_i RR^T)}}$$

$$\frac{\partial \mathcal{L}}{\partial R} = -GR + \Omega_2 R = 0,$$

$$G = \frac{W_i \hat{X}_i^T LL^T \hat{X}_i}{2\sqrt{\mathrm{Tr}(\hat{X}_i^T \hat{X}_i - \hat{X}_i^T LL^T \hat{X}_i RR^T)}}.$$

4. Experimental Results

➢ Experiment 1. Image Classification

Dataset

- MNIST dataset, 60000 for training and 10000 for testing,
- Randomly select {400,600,800} samples per digit for training.
- Randomly choose 5% of the training samples and weight them by a magnitude.
- Choose all the testing image for testing.

4. Experimental Results

➢ Experiment 1. Image Classification

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

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

4. Experimental Results

- Experiment 1. Image Classification

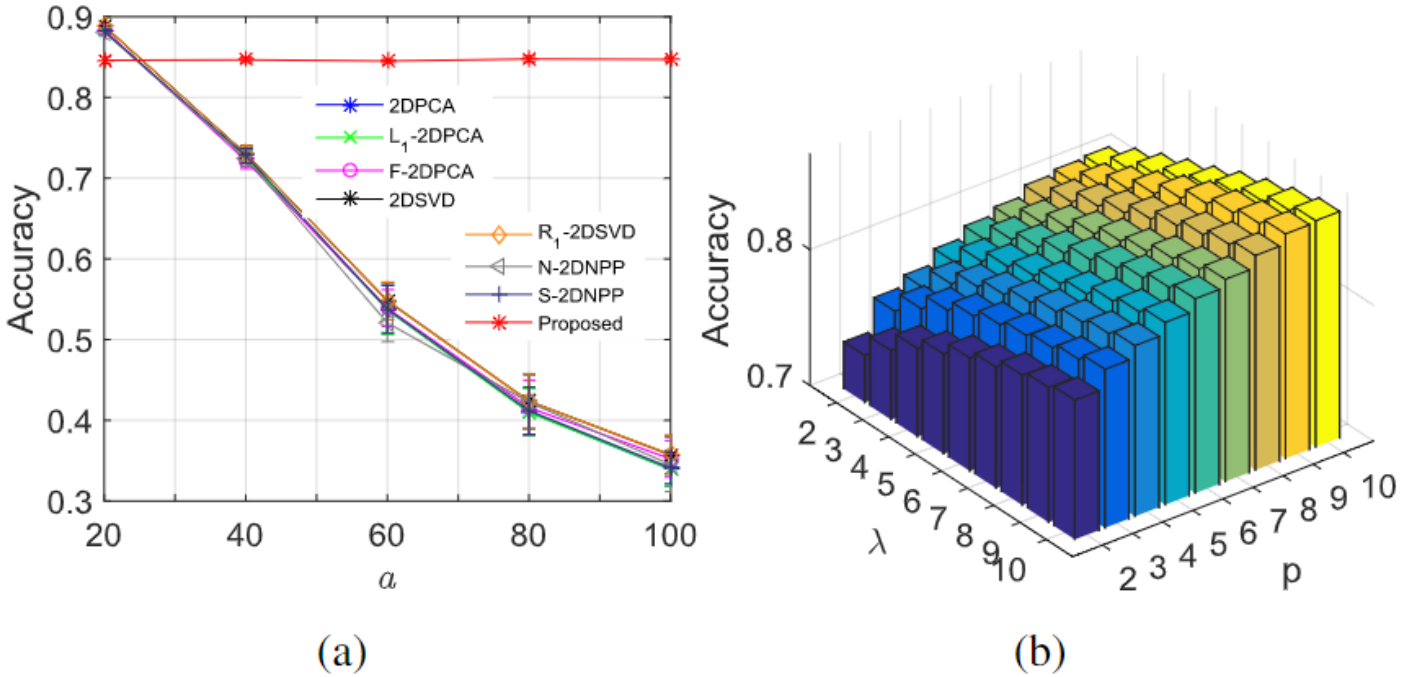


Fig. 1. AC on the MNIST Handwritten Digit Dataset. (a) AC of all the algorithms with changing magnitude of outliers; (b) AC of the proposed algorithm with different λ and p.

4. Experimental Results

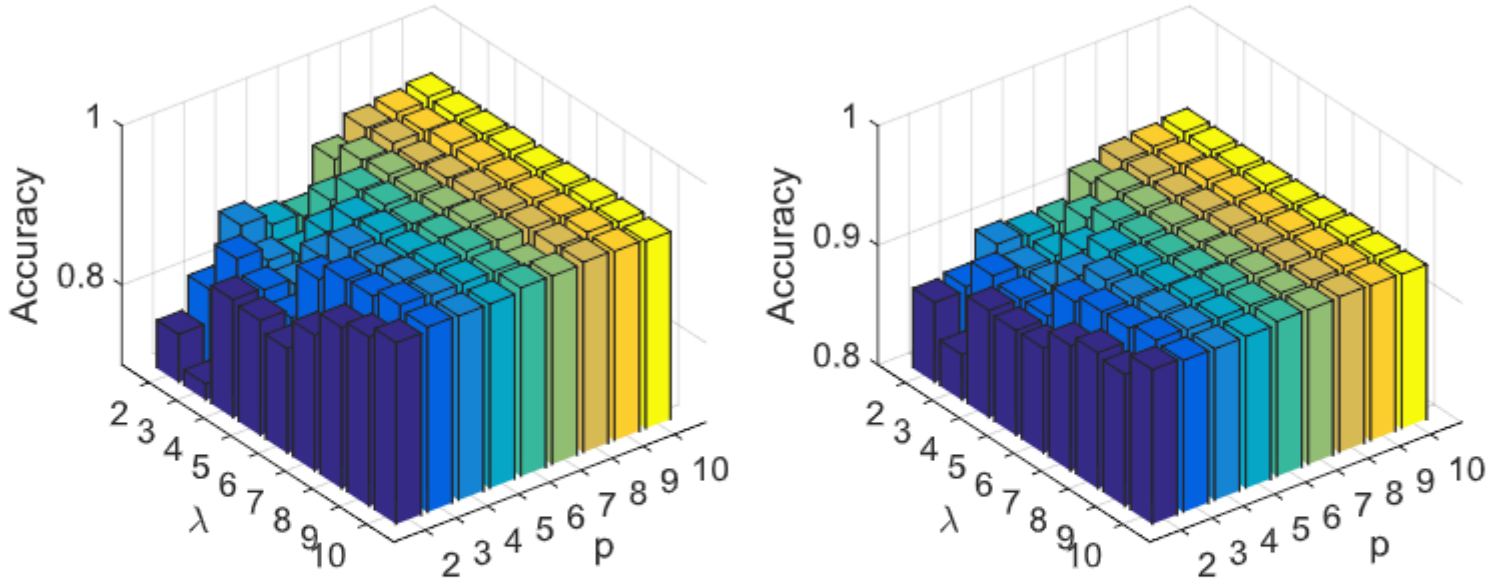
➢ Experiment 2. Image Clustering

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

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

4. Experimental Results

- Experiment 2. Image Clustering



(a) (b)

Fig. 2. AC with different λ and p on the ORL dataset. (a) AC; (b) NMI.

5. Conclusion

Conclusion

- The proposed robust 2DSVD method is robust to outliers
- Can handle non-centered data and update data mean during optimization
- Preserve the rotational invariant of the original 2DSVD methods
- Easy to be extended to higher order tensor case
- Better performance on image classification and clustering

The end of presentation
Thank you

Email: lena.zhang@griffith.edu.au