

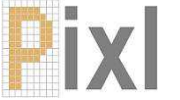


# Separable Principle Component Analysis For Image Classification

Yongxin Taylor Xi and Peter J. Ramadge

Department of Electrical Engineering, Princeton University, NJ 08544

Email: [yxi@princeton.edu](mailto:yxi@princeton.edu), [ramadge@princeton.edu](mailto:ramadge@princeton.edu)



## Introduction

Classification:

Let  $\mathcal{X}$  denote a linear space and  $\mathcal{Y}$  denote a finite set of labels.  
Given a set of training examples  $\{(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}, k = 1, \dots, N\}$   
design a classifier  $h: \mathcal{X} \rightarrow \mathcal{Y}$  that 'best' predicts labels

Dimensionality reduction:  $Q: \mathcal{X} = \mathbb{R}^s \rightarrow \mathbb{R}^d$

High dimensional  $\mathcal{X}$  E.g. grey scale  $m \times n$  face images

PCA approach:

Use the training data to determine a linear projection into a lower dimensional space, then the label information is used to design a classifier.

$$D = [x_1, x_2, \dots, x_N] \quad DD^T = \sum_{k=1}^N x_k x_k^T \in \mathbb{R}^{s \times s}, s = m \times n$$

PCA: Eigen decomposition of  $DD^T$

SVD: Singular value decomposition of  $D$

Let  $w_j$  denote the  $j$ th eigenvector ordered by eigenvalue, largest to smallest. Then  $Q = [w_1 w_2 \dots w_d]$

Limitation:

Time and space complexity of PCA/SVD:

With  $N \ll mn$ , the time cost of SVD is  $O(mnN^2)$  and the space required is  $O(mnN)$

## Recent Matrix-based PCA Methods

2D PCA: J. Yang, D. Zhang et al. 2004

$$A_k \in \mathbb{R}^{m \times n}, k = 1, \dots, N$$

$$R = \sum_{k=1}^N A_k^T A_k = \sum_{k=1}^N \sum_{i=1}^m r_i^k r_i^{kT}$$

$R$  is the scatter matrix of all rows over all training images.

Form  $V_q \in \mathbb{R}^{n \times q}$  by the first  $q$  eigenvectors.

Dimension reduction:  $\hat{A}_k = A_k V_q \in \mathbb{R}^{m \times q}$

The same procedure can be applied to columns.

$$C = \sum_{k=1}^N A_k A_k^T = \sum_{k=1}^N \sum_{j=1}^m c_j^k c_j^{kT}$$

$$\hat{A}_k = U_p^T A_k \in \mathbb{R}^{p \times n}$$

BDPCA: W. Zuo et al. 2005

$$\text{Bidirectional PCA: } \hat{A}_k = U_p^T A_k V_q$$

GLRAM: J. Ye, 2005

Generalized low rank approximations of matrices:

$$\min_{B_k, U, V} \sum_{k=1}^N \|A_k - U^T B_k V\|_F^2$$

2DPCA, BD-PCA and GLRAM:

Reported computational efficiency and empirical evidence of robust performance in image classification

## Separable PCA

BDPCA is a separable Orthonormal image transformation.

Let  $V = [V_q \tilde{V}], U = [U_p \tilde{U}]$ , then  $\{W_{i,j} = u_i v_j^T\}$  is a **separable ON basis** for  $\mathbb{R}^{m \times n}$ .

What is the optimal (in the PCA sense) projection onto a separable basis?

We pose the **Separable PCA (SPCA) Problem**:

$$\max_{\substack{U \in \mathbb{R}^{m \times p} \\ V \in \mathbb{R}^{n \times q}}} T(U, V) = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^N \langle A_k, u_i v_j^T \rangle^2$$

$$\text{subj. to: } U^T U = I_p \text{ and } V^T V = I_q$$

Simple algebraic rearrangement yields:

$$T(U, V) = \text{trace} \left( V^T \left( \sum_{k=1}^N A_k^T U U^T A_k \right) V \right) = \text{trace} \left( U^T \left( \sum_{k=1}^N A_k V V^T A_k^T \right) U \right)$$

Continuity of  $T$  and compactness of the product manifold ensures a solution.

Solution not unique: If  $(U^*, V^*)$  is a solution, then so are  $(U^* Q_1, V^* Q_2)$ ,  $Q_1 \in O^{p \times p}$ ,  $Q_2 \in O^{q \times q}$

## SPCA unifies 2DPCA, BDPCA and GLRAM

Theorem 1 gives an upper bound on  $T(U, V)$ , and provides the conditions for the bound to be tight.

Theorem 1 also provides a rule of thumb of choosing  $p$  and  $q$ .

**Theorem 1.** Let  $U \in \mathbb{R}^{m \times p}$  and  $V \in \mathbb{R}^{n \times q}$  have ON columns and  $\mathcal{U} = \mathcal{R}(U)$  and  $\mathcal{V} = \mathcal{R}(V)$ . Then

$$T(U, V) \leq \min \left\{ \sum_{j=1}^p \sigma_j(C), \sum_{j=1}^q \sigma_j(R) \right\} \quad (9)$$

with equality if either: the columns of  $U$  are the first  $p$  principal eigenvectors of  $C$  and for each  $k$ ,  $A_k^T \mathcal{U} \subseteq \mathcal{V}$ ; or the columns of  $V$  are the first  $q$  principal eigenvectors of  $R$  and for each  $k$ ,  $A_k \mathcal{V} \subseteq \mathcal{U}$ .

**Corollary 1.1 Optimality of BD-PCA (Easily checkable)**

Let  $\mathcal{V}_q = \mathcal{R}(V_q)$  and  $\mathcal{U}_p = \mathcal{R}(U_p)$

If  $\forall k, A_k \mathcal{V}_q \subseteq \mathcal{U}_p$  or  $\forall k, A_k^T \mathcal{U}_p \subseteq \mathcal{V}_q$  then **BD-PCA solves SPCA**.

**Corollary 1.2** If  $p=m$  or  $q=n$ , **BD-PCA solves SPCA**.

**Corollary 1.3** 2DPCA solves a special case of BD-PCA.

2DPCA is obtained from BD-PCA by the restriction  $p=m$  and  $U_m = I_m$

**Theorem 1.4 GLRAM is equivalent to SPCA. [Ye, 2005]**

Let  $U \in \mathbb{R}^{m \times p}$ ,  $V \in \mathbb{R}^{n \times q}$ ,  $B_k \in \mathbb{R}^{p \times q}$

GLRAM solves  $\min_{U, V, B_k} \sum_k \|A_k - U B_k V^T\|_F^2$

**Solving SPCA** (nontrivial):

- GLRAM algorithm: iteratively update  $U$  and  $V$  until local optimum.

- NGLRAM algorithm: update only once from BD-PCA solutions.

## Experiments

**Databases** - Handwritten Digits database (UCI repository)

- YALE face database (resized to 60x50)

- ORL face database

**Algorithms:** PCA, SPCA (GLRAM), 2DPCA, BD-PCA and NGLRAM

Each experiment is repeated 20 times with random selection of the training set.

**Training:** 100 per digit for Digits, 6 per person for YALE, 5 per person for ORL

**Testing:** root mean square reconstruction error (RMSRE)

test image misclassification rate

SPCA is computed using the GLRAM algorithm

(stopping criterion:  $\Delta \text{RMSRE} < 10^{-6}$ )

Data Set	$m \times n$	$N$	$mn/N$	$p_{\max}^{\text{PCA}}$
Digits	$32 \times 32$	1000	1.028	31
Yale	$60 \times 50$	90	33.3	9
ORL	$112 \times 92$	200	206	14

**Observations:**

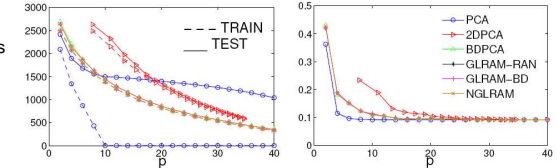
1. GLRAM algorithm was robust to variations in initial conditions.

2. Both BD-PCA and NGLRAM provided very good approximate SPCA solutions.

3. BD-PCA is faster to compute and requires less memory than PCA.

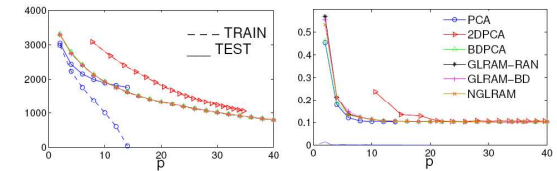
4. For large values of  $mn/N$  SPCA was more resistant to overfitting.

5. However, the reduced over fitting of SPCA did not improve classification performance.



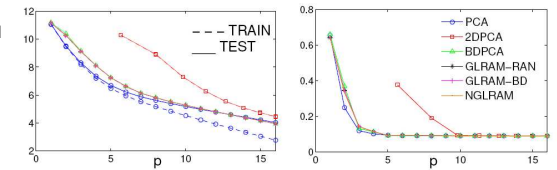
(a) YALE: Reconst. error

(b) YALE: Test error rate



(c) ORL: Reconst. error

(d) ORL: Test error rate



(e) DIGITS: Reconst. error

(f) DIGITS: Test error rate

## Conclusion

1. SPCA unifies recently proposed matrix-based dimension reduction methods.
2. Robustness to overfitting of SPCA is an advantage in image approximation applications, but not necessarily classification accuracy.
3. Fast algorithms e.g. BD-PCA, NGLRAM give very good approximate SPCA solutions.
4. For resource constrained applications, SPCA methods save much memory.

## References [See the paper]