Journal of Data Science 24 (4), 734–751
October 2026
DOI: 10.6339/24-JDS1158
Statistical Data Science

SPA: Signflip Parallel Analysis to Optimize the Number of Principal Components in Two-dimensional PCA

Zhaoyuan Li[1] and Yiling Kuang[2]
[1]*The Chinese University of Hong Kong, Shenzhen, China*
[2]*The Chinese University of Hong Kong, China*

Abstract

Yang et al. (2004) developed the two-dimensional principal component analysis (2DPCA) for image representation and recognition, widely used in different fields, including face recognition, biometrics recognition, cancer diagnosis, tumor classification, and others. 2DPCA has been proven to perform better and computationally more efficiently than traditional principal component analysis (PCA). However, some theoretical properties of 2DPCA are still unknown, including determining the number of principal components (PCs) in the training set, which is the critical step in applying 2DPCA. Without rigorous criteria for determining the number of PCs hampers the generalization of the application of 2DPCA. Given this issue, we propose a new method based on parallel analysis to determine the number of PCs in 2DPCA with statistical justification. Several image classification experiments demonstrate that the proposed method compares favourably to other state-of-the-art approaches regarding recognition accuracy and storage requirement, with a low computational cost.

Keywords *2DPCA; feature extraction; image analysis*

1 Introduction

PCA is a classical feature extraction and data representation technique widely used in pattern recognition and computer vision. One of the first successful methods for image-based face recognition was developed by Turk and Pentland (1991), known as Eigenfaces. Although it is natural to transform two-dimensional (2D) face image matrices into one-dimensional (1D) long image vectors to use standard PCA, the resulting image vectors of faces usually lead to a high-dimensional image vector space, where the sample covariance matrix computed using a relatively small number of training samples is no longer a reasonable estimation of the population covariance, according to the recent theory development of high-dimensional statistics (Bai and Silverstein, 2010). Meanwhile, it is often the case that the columns and rows of a matrix represent different sets of information that are closely interrelated in a very structural way. Yang et al. (2004) developed the 2DPCA method that maintains and utilizes the matrix structure to achieve more significant dimension reduction.

The 2DPCA method remains a prevalent and successful technique. Applications abound in face recognition (Ejaz et al., 2019), cancer diagnosis (Dhahri et al., 2019), human activity classification (Steven Eyobu and Han, 2018), remote sensing image classification (Uddin et al., 2021), medical multi-modal retrieval (Zeng et al., 2024), cancellable biometrics (Yang et al.,

*Corresponding author. Email: lizhaoyuan@cuhk.edu.cn.

© 2026 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received July 23, 2024; Accepted October 20, 2024

2024), space-time-coding digital metasurface (Wang et al., 2024), brain cancer classification (Gumaei et al., 2019), image colourization (Wan et al., 2020) and many others. In addition, based on 2DPCA, many methods have been developed, such as optimal mean 2DPCA (Wang et al., 2017) and novel folded-PCA (Zabalza et al., 2014). However, some theoretical properties of 2DPCA are still unknown, including determining the number of PCs in the training model, which is the critical step in applying 2DPCA. Estimating how many PCs to keep is well known to impact downstream data analyses significantly. Without statistically rigorous criteria for determining the number hampers the generalization of the application of 2DPCA.

In applying PCA, selecting the number of PCs to keep is one of the most critical problems. However, existing methods, such as the scree plot, likelihood ratio, permutation parallel analysis, and eigenvalue-based methods, were developed for measurements of p features over a set of n samples (referred as to 1D vector data), do not have statistical guarantees in 2D matrix data. Different from independent and identically distribution assumptions on 1D vector data, the noise of 2D matrix data could be heterogeneous (each noise entry can have a different distribution). Moreover, many existing methods select the number of PCs subject to subjective judgment (see Section 2). For 2DPCA, Yang et al. (2004) proposed to use the top recognition accuracy on test data to determine the best choices of the number of PCs in the training model. However, this method often gives different choices for different test data, as seen in the experiments in Section 4. On the one hand, the amount and quality of test data would significantly affect the choice of the number of PCs. On the other hand, the best choices of the number of PCs for the training model depend on the test data and are not known beforehand in a real problem, which hampers the generalization of the application of 2DPCA.

This paper aims to develop a novel method to determine the number of PCs in 2DPCA with rigorous statistical justification. To tackle this issue, our conceptual idea is to find a "null" data containing only the noise level of the original data, without signals, and thus naturally determine the number of PCs by comparing the original data and the "null" counterpart. Guided by this insight, we build a new signflip parallel analysis (SPA) algorithm. A "null" copy of data is generated by randomly, independently, and uniformly flipping the signs of the data entries. By conducting this procedure many times, the empirical behaviour of the noise level of the original data is obtained. Then, the number of PCs can be determined by comparing each eigenvalue of the original sample covariance matrix to a percentile of the constructed empirical distribution with a statistical significance level. Therefore, this proposed method is statistically guaranteed instead of subjective judgment. Extensive experiments show that the proposed method performs very well and can generalize the application of 2DPCA.

The main contributions of this paper are summarized as follows. (1) The proposed SPA algorithm provides the best choice of the number of PCs in 2DPCA in training data with a statistical significance level. (2) Based on the SPA algorithm, 2DPCA is generalized as the training model is determined beforehand in a real problem. (3) The SPA algorithm achieves state-of-the-art results regarding both recognition accuracy and storage requirements on three public image datasets, from the Olivetti Research Laboratory (ORL) database, the Face Recognition Technology (FERET) database, and the extended Yale Face B database, respectively. In addition, the computational complexity of the SPA algorithm is low.

The remainder of this paper is organized as follows. Section 2 reviews classical methods of determining the number of PCs in PCA. The idea of the proposed SPA algorithm is described in Section 3. Section 4 presents experimental results for the ORL, FERET, and extended Yale Face B face datasets to demonstrate the effectiveness of the SPA algorithm and the generalization of 2DPCA. Finally, conclusions are presented in Section 5.

2 Related Work

As discovering latent low-dimensional phenomena in large and messy datasets is one of the central challenges faced in modern data analysis, much work has gone into developing many methods. Indeed, many more can be discussed in detail here, so we give a brief high-level overview instead.

One of the classical and standard methods is the scree plot (Cattell and Vogelmann, 1977), i.e., Cattell's scree plot. Construct the so-called scree plot of the descending-order eigenvalue ℓ_i of a matrix on the vertical axis versus i on the horizontal axis with equal intervals for $i = 1, \ldots, p$, and join the points into a decreasing polygon. A "clean-cut" where the polygon "levels off" so that the first few eigenvalues seem to be far apart from the others. The elbow joint of the scree plot must be located, and the number of PCs is the level of i, which is just on the left-hand side of the elbow joint. However, the elbow of a scree plot is observed subjectively and sometimes is unclear to be observed. There is no statistically significant threshold to determine the elbow.

Based on the magnitudes of eigenvalues, another classical method is to include the components such that the cumulative proportion of the total variance explained is just more than a threshold value, say 80%, i.e., if $\sum_{i=1}^{q} \ell_i / \sum_{i=1}^{p} \ell_i > 0.8$ (eigenvalues $\{\ell_i\}$ are in descending order), then q PCs can be kept (Hair et al., 1986). This method is referred to as *total variance*. However, choosing the cumulative proportion of the total variance is arbitrary and subjective. Moreover, for a large dimensional dataset, the number of principal components selected could still be very large, even if the cumulative proportion of the total variance is low (Bai and Silverstein, 2010).

Another more rigorous way is to use the $(1-\alpha)100\%$ upper confidence limit of the parametric function $g(\ell) = \sum_{i=q+1}^{p} \ell_i / \sum_{i=1}^{p} \ell_i$ (eigenvalues $\{\ell_i\}$ are in descending order), which measures the relative importance of last $(p-q)$ eigenvalues to all p eigenvalues,

$$g(\ell) + z_\alpha \frac{\sqrt{2}}{\sqrt{n-1}\sum_{i=1}^{p} \ell_i} \sqrt{g^2(\ell) \sum_{i=1}^{q} \ell_i^2 + \big(1 - g(\ell)\big)^2 \sum_{i=q+1}^{p} \ell_i^2},$$

where z_α is the right tail cut-off point of the standard normal distribution with probability α; e.g., $z_{0.05} = 1.645$, and n is the sample size to calculate the matrix. If the upper confidence limit is sufficiently small, say less than a threshold proportion of 0.1, the eigenvalues $\ell_{q+1}, \ell_{q+2}, \ldots, \ell_p$ are too small, and they are not helpful to explain the variation of data. Thus, one can retain the first q PCs only. On the other hand, if the upper confidence limit is larger than the threshold, ℓ_{q+1} is still useful to explain the variation of data, and one should keep the $(q+1)$ PCs. This is referred to as UCL method. However, similar with the total variance, the threshold for UCL also has to be chosen subjectively.

Recently, some methods have been proposed based on the relationship of adjacent eigenvalues to determine the number of PCs, including the difference between consecutive eigenvalues (DBCEigen) and the ratio of consecutive eigenvalues (RCEigen). Onatski (2010) proposed to use the differences between adjacent eigenvalues, and Lam and Yao (2012), Wang (2012), and Ahn and Horenstein (2013) analogously proposed using the ratio. For descending-order eigenvalues $\{\ell_i\}$, the DBCEigen and RCEigen are defined, respectively, as

$$d_i = \ell_i - \ell_{i+1}, \quad i = 1, \ldots, p-1,$$
$$r_i = \ell_i / \ell_{i+1}, \quad i = 1, \ldots, p-1.$$

The eigenvalues are selected to be kept until the above difference d_i or ratio r_i are less than a threshold. However, existing works constructed thresholds for 1D vector data under some

unique structures are unsuitable for 2D matrix-valued data. Therefore, the thresholds are selected arbitrarily when these methods are used in 2DPCA.

Another popular and practical method is permutation parallel analysis proposed by Horn (1965) and Buja and Eyuboglu (1992), and there is a large amount of evidence that parallel analysis is one of the most accurate methods for determining the number of PCs in PCA (Owen and Wang, 2016). However, this method works well for homogeneous noise and can dramatically degrade when noise is heterogeneous. To solve this problem, Hong et al. (2020) proposed the signflip parallel analysis method for large-dimensional data with heterogeneous noise. These methods are also designed for 1D vector data.

Yang et al. (2004) proposed to use the top recognition accuracy on test data to determine the best choices of the number of PCs in the training data. For simplicity, this method is referred to as Top-RA. However, this method often gives different choices given different test data, as seen in the experiments in Section 4. On the one hand, the amount and quality of test data would significantly affect the choice of the number of PCs. On the other hand, using test data to determine the model's rank in training data hampers the generalization of the application of 2DPCA.

In this paper, we introduce signflip parallel analysis to matrix-valued data and develop a new algorithm to determine the number of PCs in 2DPCA.

3 Signflip Parallel Analysis

Let X denote an image matrix with dimension $n \times p$. The image covariance matrix is defined as

$$Cov(X) = \mathbb{E}\big[(X - \mathbb{E}X)'(X - \mathbb{E}X)\big].$$

Suppose that there are m training samples, the jth $n \times p$ matrix is X_j $(j = 1, 2, \ldots, m)$, and the average matrix of all samples is $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_j$. Then the sample image covariance matrix is

$$G = \frac{1}{m}\sum_{j=1}^{m}\big(X_j - \bar{X}\big)'\big(X_j - \bar{X}\big). \tag{1}$$

It is easy to verify that G is a $p \times p$ non-negative definite matrix. 2DPCA is based on this matrix G to achieve feature extraction and classification.

Given images data $\{X_j\}_{j=1}^{m}$ and the corresponding sample image covariance matrix G, our goal is to determine the number of PCs of matrix G retained for 2DPCA. A SPA algorithm is proposed based on comparing the eigenvalues of G to those of "empirical null" data generated by randomly, independently, and uniformly flipping the signs of the data matrix entries. The selected rank is the number of leading data eigenvalues that rise above their signflipped analogs, where the comparison is made sequentially, starting from the top eigenvalue and stopping at the first failure. The algorithm is described in Algorithm 1.

In Algorithm 1, without changing the noise level, steps 4–5 generate one "parallel" copy of the sample image covariance G every time, which can be treated as the "null" pure noise analog. And this procedure is repeated T times, so that T "parallel" copies $\tilde{G}^{(t)}$ of the sample image covariance G are obtained. The key idea is that we expect components rising above the noise to produce data eigenvalues above their pure-noise analogs. Thus, the eigenvalues $\{\tilde{\lambda}_1^{(t)} \geqslant \cdots \geqslant \tilde{\lambda}_p^{(t)}, t = 1, \ldots, T\}$ of the signflipped matrices $\tilde{G}^{(t)}$, $t = 1, \ldots, T$, can be used to

Algorithm 1 SPA: select the number of PCs in 2DPCA

Input: Training data $X_j \in \mathbb{R}^{n\times p}$, $j = 1, \ldots, m$, percentile α, number of trials T

Output: Selected number of PCs $\hat{q}$

1: $G \leftarrow$ sample image covariance matrix

2: $\lambda_1 \geqslant \cdots \geqslant \lambda_p \leftarrow$ eigenvalues of G

3: **for** $t \leftarrow 1$ **to** T **do**

4: Randomly signflip entries of G to calculate $\tilde{G}^{(t)}$: form $R_t \circ G$ where

$$R_t(ij) \overset{i.i.d.}{\sim} \begin{cases} +1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2, \end{cases}$$

i.e. $R_t \in \mathbb{R}^{n\times p}$ has independent identically distributed Rademacher entries;

5: $\tilde{\lambda}_1^{(t)} \geqslant \cdots \geqslant \tilde{\lambda}_p^{(t)} \leftarrow$ eigenvalues of $\tilde{G}^{(t)}$;

6: **end for**

7: $\hat{q} \leftarrow$ first q for which either

$$\lambda_{a+1} \leqslant \alpha\text{-percentile of } \left\{\tilde{\lambda}_1^{(1)}, \ldots, \tilde{\lambda}_1^{(T)}\right\}, \text{upper-edge}$$

or

$$\lambda_{q+1} \leqslant \alpha\text{-percentile of } \left\{\tilde{\lambda}_{q+1}^{(1)}, \ldots, \tilde{\lambda}_{q+1}^{(T)}\right\}, \text{pairwise}$$

i.e., q is the number of leading eigenvalues above the α-percentile of their signflipped analogs, where "pairwise" and "upper-edge" are two choices for comparison.

obtain the empirical distribution of the noise eigenvalues of G. Correspondingly, we can find a threshold from the empirical distribution with a statistical significance level $\alpha = (0.1/0.05/0.01)$. In the SPA algorithm, there are two ways to select the threshold:

1. (upper-edge) The upper-edge comparison compares all data eigenvalues against (the α percentile of) only the largest (first) signflipped eigenvalues;
2. (pairwise) The pairwise comparison selects the number of leading eigenvalues above the α percentile of their signflipped analogs.

The upper-edge comparison never selects more principal components than the pairwise comparison, making it more conservative (see experiments in Section 4). Moreover, the upper-edge comparison has the benefit of only requiring us to calculate and store the first eigenvalue. The two selection rules are essentially asymptotically equivalent and agree in many settings.

For the number of trials T in the Algorithm 1, generally, it should be as large as possible to retain a stable and accurate result. However, many repetitions lead to time-consuming computation, especially for large dimensional data. To our best knowledge, a suitable number of trials is from 20 to 100. Specifically, the minimum number of trials can be 20 if $\alpha = 0.1$ or $\alpha = 0.05$ is used. But the number of trials should be larger (say 100 or more) when $\alpha = 0.01$ is used.

Based on this SPA algorithm, we can determine the model's rank in the training data beforehand in a real problem, thus achieving the generalization of the application of 2DPCA.

4 Numerical Studies

We illustrate the performance of the proposed SPA algorithm in image classification tasks using three publicly available datasets and compare SPA with the state-of-the-art existing methods mentioned in Section 2, including scree plot, total variance, UCL, DBCEigen, RCEigen and Top-RA. SPA-u denotes SPA with upper-edge comparison, SPA-p denotes SPA with pairwise comparison.

The procedure of 2DPCA in image classification is as follows:

1. Given training data, the sample image covariance G in (1) is calculated;
2. The number k of PCs of G is determined;
3. The corresponding k eigenvectors of G, $u_1, \ldots, u_k$, are used as the optimal projection vectors;
4. Calculate feature image of the image sample: $Y_j = X_j \cdot P$, $P = [u_1 \cdots u_k]_{p\times k}$, for $j = 1, \ldots, m$;
5. For a new image X, a nearest neighbour classifier is used for classification:

$$Y = X[u_1 \cdots u_k],$$
$$j = \arg\min_{j'} d(Y, Y_{j'}), \quad j' = 1, \ldots, m,$$

where $d(A, B)$ is the Euclidean distance between A and B, then X is classified to the class of X_j.

So, all the methods of determining the number of PCs are applied in the second step. We use three metrics, stability, average testing accuracy (ATA), and an overall score of 2DPCA classification, to quantitatively evaluate the performance of these methods, where the stability and overall score are defined as:

$$\text{stability} = \text{standard deviation (\# of selected PCs)}, \tag{2}$$

$$\text{overall score} = \text{mean}\left(\frac{\text{testing accuracy}}{\text{\# of PCs}}\right). \tag{3}$$

Compared to PCA, one disadvantage of 2DPCA is that more coefficients are needed to represent an image (Yang et al., 2004). The more PCs are retained in 2DPCA, the more coefficients are used; thus, more storage is required. Therefore, the selected number of PCs is divided by corresponding testing accuracy in the overall score to evaluate the overall performance of different methods.

4.1 Experiments on the ORL Database

The ORL database (http://cam-orl.co.uk/facedatabase.html) contains images from 40 individuals, i.e., 40 classes, each providing 10 different images. For some individuals, the images were taken at different times. The lighting, facial expressions (open or closed eyes, smiling or not smiling), and facial details (glasses or no glasses) also vary. All the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The size of each image is 92×112 pixels, with 256 grey levels per pixel. Five sample images of one person from the ORL database are shown in Figure 1.

The ORL dataset was used to evaluate the performance of SPA under conditions where the pose is varied. We use the first one to five ($k = 1, 2, 3, 4, 5$) image samples per class for training and the remaining images for testing. Thus, the total number of training samples is 40, 80, 120, 160, and 200, respectively, and the corresponding number of testing samples is 360, 320,

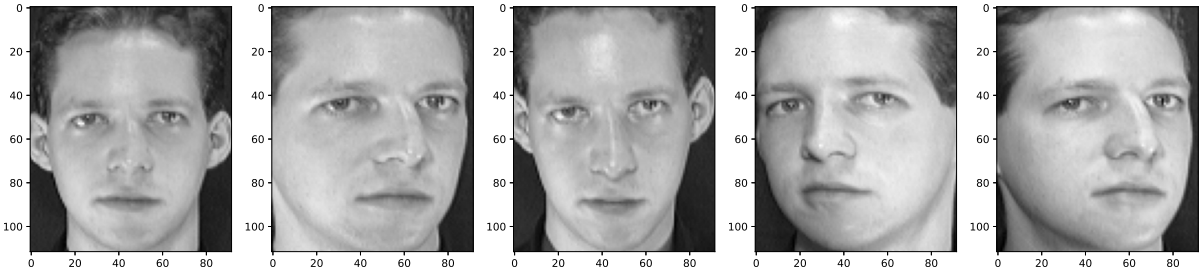


Figure 1: Five sample images of one person in the ORL face database.

280, 240, and 200, respectively. The image covariance matrix G is calculated from the training samples for each case.

First, the results of SPA are presented in Figure 2. The number of trials is $T = 100$. The five plots on the left panel show the first 60 largest eigenvalues of G of training samples (original data), along with their signflipped analogs (permuted data) for the five cases, $k = 1, 2, 3, 4, 5$, respectively. Notice that the magnitude of the eigenvalues is enormous for the first several, but quickly converges to zero. This shows that the information of an image is concentrated on its first small number of component vectors. For all cases, $k = 1, 2, 3, 4, 5$, SPA with upper-edge comparison selects the first three largest eigenvalues in the training data, and SPA with pairwise comparison selects the first four largest eigenvalues. The selected number of PCs keep the same, demonstrating the stability of the SPA algorithm.

To illustrate the Top-RA method, the five plots on the right panel of Figure 2 present the classification accuracy of 2DPCA on the test data against different numbers of eigenvalues kept in the training set. As the number of eigenvalues kept in the training set increases, the testing classification accuracy increases first and then decreases quickly due to more noise being contained when a relatively large number of eigenvalues is kept. For cases, $k = 1, 2, 3, 4, 5$, the number of eigenvalues with top classification accuracy on the test data are 5, 4, 6, 6, and 7, respectively, which implies that the sample size in the test data influences the selected number of PCs. Therefore, it is not a stable way to use the Top-RA method to decide the number of PCs.

The scree plot is just the line of original data in each plot on the left panel of Figure 2. As mentioned before, we must subjectively decide the elbow of a scree plot. So 3, 4, and 5 are all possible choices. By setting the threshold as 0.8, the total variance method chooses 8, 8, 7, 7, and 7 numbers of PCs for five cases, respectively. By setting the threshold as 0.1 and the significance level as $\alpha = 0.05$, the UCL method selects 15, 16, 16, 15, and 16 numbers of PCs for five cases, respectively. To present the DBCEigen and RCEigen methods, the differences and ratios of consecutive eigenvalues are plotted in Figure 3, respectively. The patterns of the five cases are similar but different, especially the fluctuating "elbows", which also casts a shadow on finding clear thresholds to determine the number of PCS. The threshold for DBCEigen is set to be 10^5, and the selected numbers of PCs are 11, 11, 11, 9, and 9 for five cases. The threshold for RCEigen is set to be 1.25, and the selected numbers of PCs are 5, 4, 6, 6, and 7, respectively.

For comparison, Table 1 summarizes the above results of these methods and the metrics of their performances in terms of stability, ATA and over score. In terms of stability, SPA is the best. By design, the Top-RA method has the highest ATA of 87.1%, which is the optimal level of accuracy of 2DPCA. The ATA of SPA is very close to this optimal accuracy. Using 4 PCs selected with pairwise comparison, SPA has 86.1% ATA, less than Top-RA by 1% only. Using

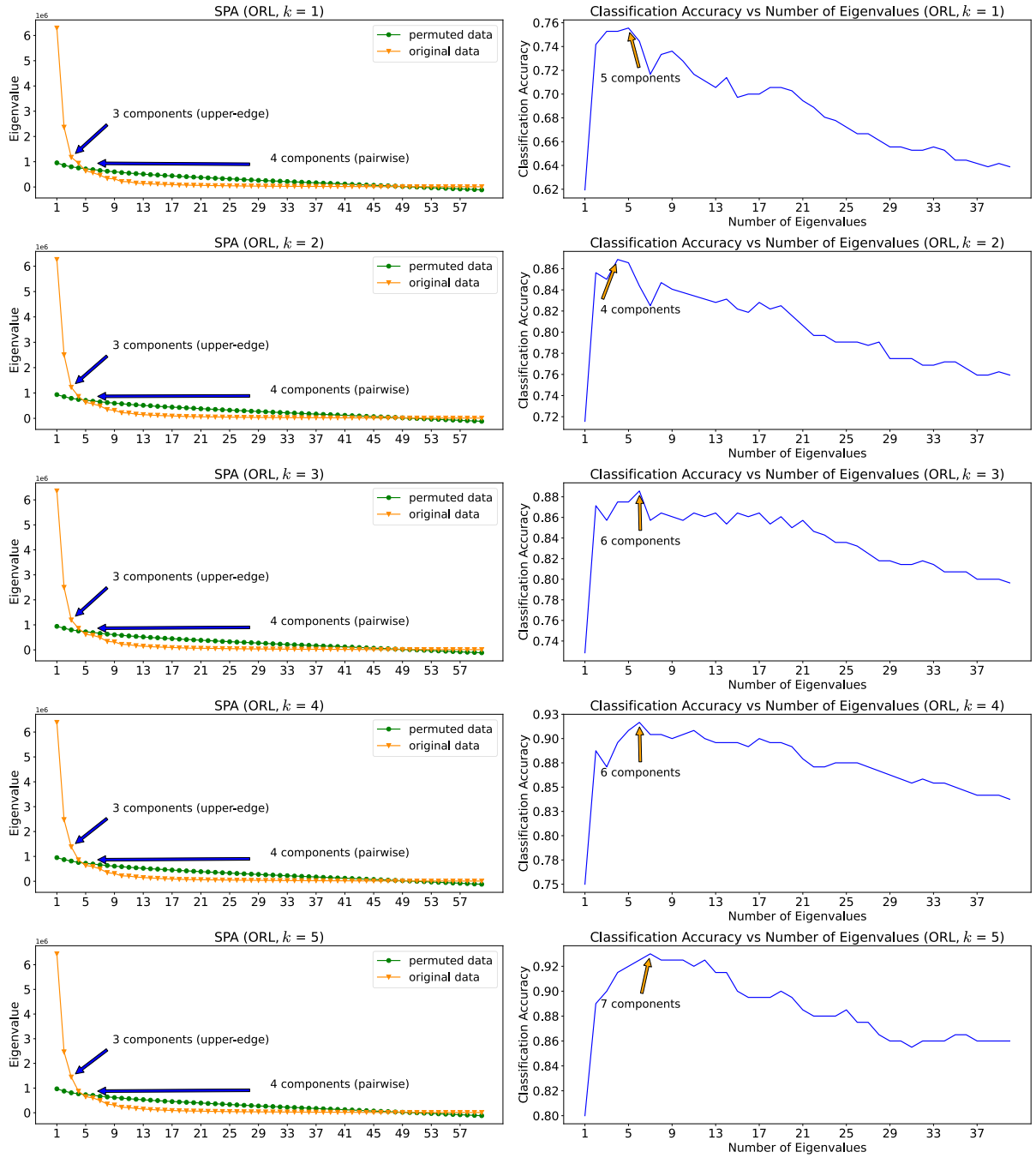


Figure 2: For the ORL database. Left panel: number of PCs selected by SPA; Right panel: classification accuracy on test data against different numbers of PCs.

3 PCs selected with upper-edge comparison, SPA has 84.6% ATA, less than Top-RA by 2.5%. Therefore, in terms of ATA, SPA with pairwise comparison is the best (except for the Top-RA). Regarding recognition accuracy and storage requirements, SPA with upper-edge comparison has the highest overall score. Compared to SPA, other methods select more PCs without improving

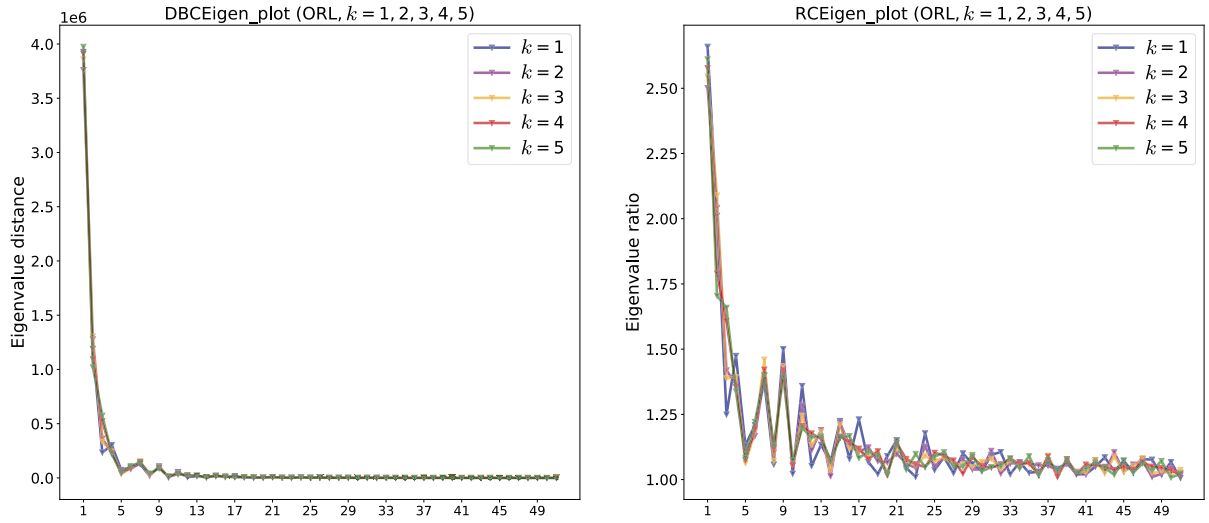


Figure 3: For the ORL database. Left: differences between consecutive eigenvalues; Right: ratios of consecutive eigenvalues.

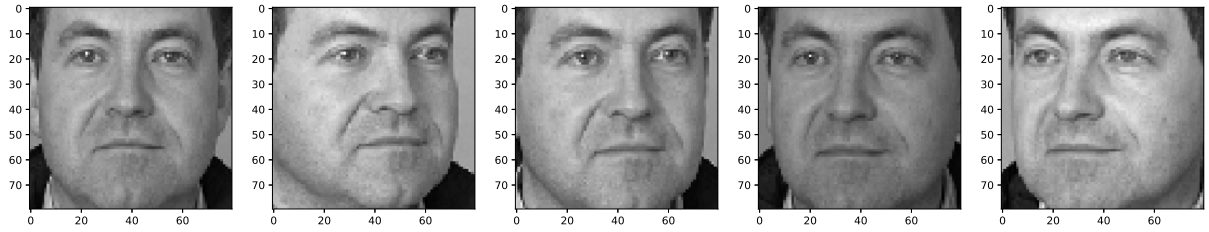


Figure 4: Sample images for one person of the FERET database.

classification accuracy. The UCL method selects the most number of PCs but leads to the lowest ATA. Total variance, DBCEigen and RCEigen also select a relatively larger number of PCs, but the corresponding ATAs are similar to that of SPA with upper-edge comparison. Thus, these methods have a low overall score. Therefore, by using SPA, 2DPCA not only can be generalized to any test but also can be close to optimal accuracy on average.

4.2 Experiment on the FERET Database

The facial images of the FERET database were collected between December 1993 and August 1996, accumulating 14,126 images on 1199 individuals and 365 duplicate sets of images taken on a different day. A subset of this database is used here, which contains the still face image of 200 individuals, each with 7 images. All these images are in tif file format, with RGB color model. To apply the 2DPCA algorithm, we transformed each image into a grayscale in a 80×80 matrix. In contrast to the ORL database, the facial expressions and lighting conditions are different, the images show most of the face (missing part of the face, e.g., forehead) and less background, and we have more classes in the FERET dataset. Five transformed samples of one subject in the FERET database are shown in Figure 4.

The FERET dataset is used to evaluate the performance of SPA under conditions where the facial expression and lighting are varied. Similarly, we use the first one to five ($k = 1, 2, 3, 4, 5$)

Table 1: Results and comparison of SPA, Scree plot, Total variance, UCL, DBCEigen, RCEigen and Top-RA in the ORL database. (4 PCs were used to calculate the three metrics for the scree plot, but the number of PCs may vary from person to person.)

	SPA-u	SPA-p	Scree	Total var.	UCL	DBCEigen	RCEigen	Top-RA
$k = 1$	3	4	3–5	8	15	9	11	5
$k = 2$	3	4	3–5	8	16	7	11	4
$k = 3$	3	4	3–5	7	16	7	11	6
$k = 4$	3	4	3–5	7	15	7	9	6
$k = 5$	3	4	3–5	7	16	7	9	7
stability	0	0	0	0.5477	0.5477	0.4472	1.0954	1.1402
ATA	84.6%	86.1%	86.1%	85.4%	83.4%	85.0%	84.8%	87.1%
overall score	28.2	21.5	21.5	11.6	5.3	11.6	8.4	16.0

image samples per class (per person) for training and the remaining images for testing. Thus, the total number of training samples is 200, 400, 600, 800, and 1000, respectively, and the corresponding number of testing samples is 1200, 1000, 800, 600, and 400, respectively. The image covariance matrix G is calculated on the training samples for each case.

First, the results of SPA are presented in Figure 5. The number of trials is $T = 50$. The five plots on the left panel show the first 40 largest eigenvalues of G of training samples (original data) and their signflipped analogs (permuted data) for five cases $k = 1, 2, 3, 4, 5$, respectively. Same with the ORL dataset, the magnitude of the eigenvalues is very large at first and then quickly converges to zero. For all cases, $k = 1, 2, 3, 4, 5$, SPA with upper-edge comparison selects the first two largest eigenvalues in the training data, and SPA with pairwise comparison also selects the first two largest eigenvalues for cases $k = 1, 2, 3, 4$, and the first three largest eigenvalues in the last case $k = 5$.

To illustrate the Top-RA method, the five plots on the right panel of Figure 5 present the classification accuracy of 2DPCA on the test data against different numbers of PCs kept in the training model. As the number of PCs kept increasing, the testing classification accuracy increases first and then decreases quickly as more noise was contained. For cases, $k = 1, 2, 3, 4, 5$, the numbers of PCs with top classification accuracy on test data are 3, 2, 2, 6, and 13, respectively, which implies that the sample size in the test data greatly affects the selected number of PCs. Therefore, Top-RA is an unstable way in this dataset.

The scree plot is the line of original data in each plot on the left panel of Figure 5. The elbows of these scree plots are easily observed, and 3 PCs are selected for all cases. By setting the threshold as 0.8, the total variance method selects 5, 6, 6, 6, and 6 numbers of PCs for five cases, respectively. By setting the threshold as 0.1 and the significance level as $\alpha = 0.05$, the UCL selects 14, 15, 16, 16, and 16 PCs for five cases, respectively. The differences and ratios of consecutive eigenvalues are plotted in Figure 6, respectively. From the left panel, it seems clear to find a threshold for DBCEigen subjectly. But there is not a clear threshold for RCEigen showed in the right panel. The threshold for DBCEigen is set to be 3×10^4, and the selected numbers of PCs are 7 for all cases. The threshold of RCEigen is set as 1.25, and the selected numbers of PCs are 7, 13, 11, 16, and 11 for five cases, respectively.

For comparison, Table 2 summarizes the above results of these methods and the metrics of their performances in terms of stability, ATA and over score. Overall, the scree plot performs

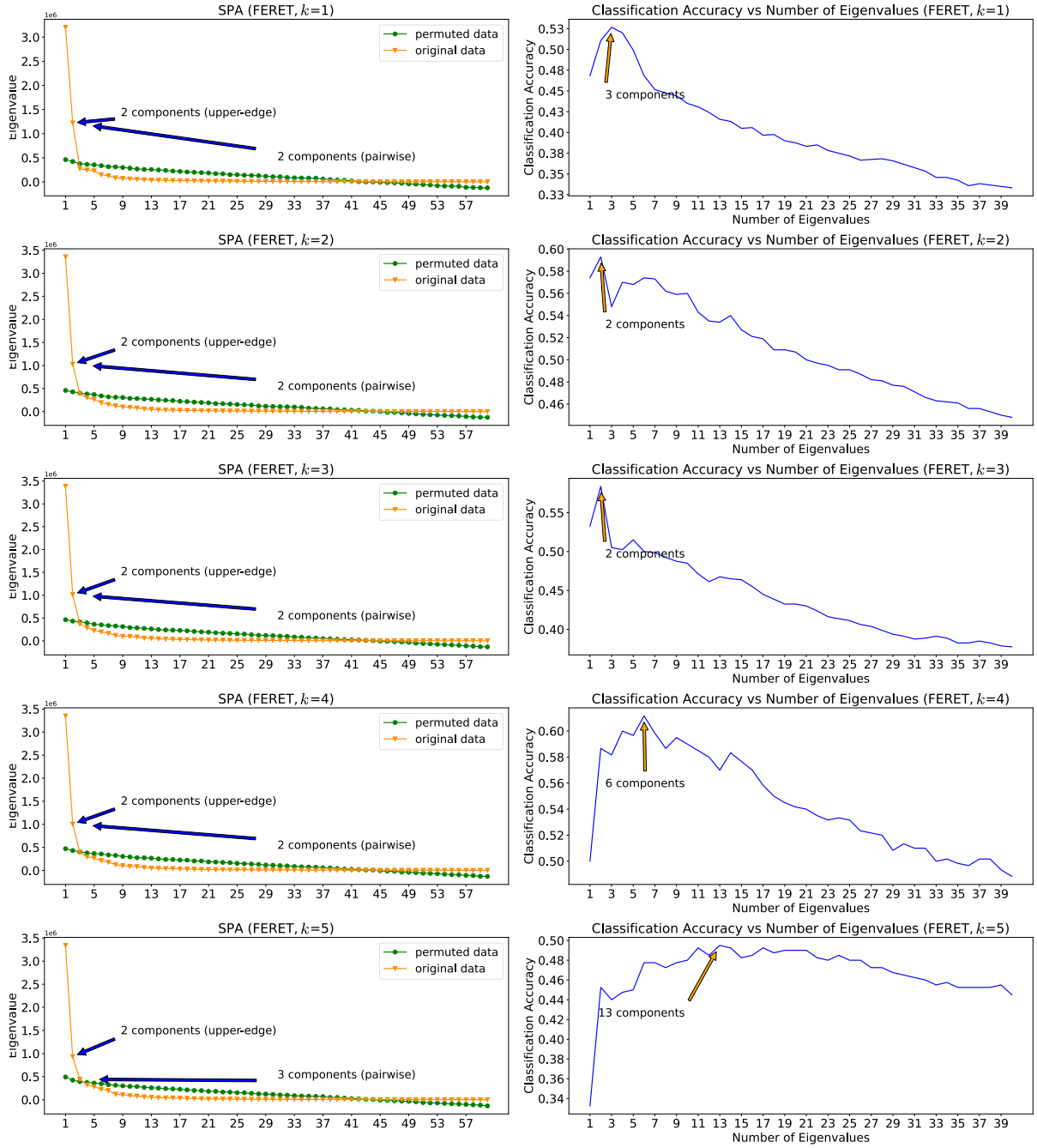


Figure 5: For the FERET database. Left panel: number of PCs selected by SPA; Right panel: classification accuracy on test data against different numbers of PCs.

close to SPA. The total variance, significance test, DBCEigen, and RCEigen methods choose much more eigenvalues with lower efficiency. In terms of stability, SPA with upper-edge comparison, scree plot and DBCEigen are the best. Given in the Top-RA, the optimal accuracy of 2DPCA for this dataset is 56.2 percent. SPA with upper-edge comparison, SPA with pairwise comparison, and scree plot have average testing accuracies close to the optimal accuracy than

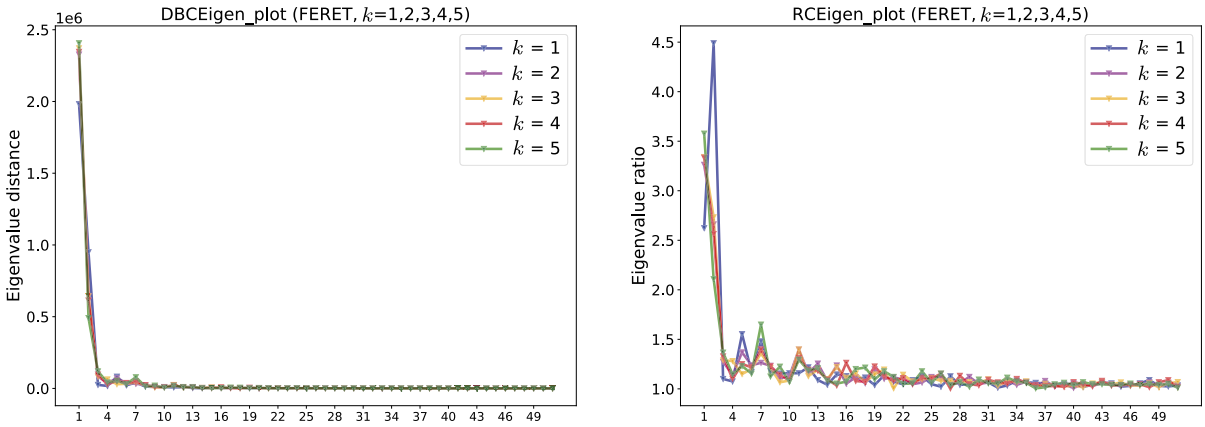


Figure 6: For the FERET database. Left: differences between consecutive eigenvalues; Right: ratios of consecutive eigenvalues.

Table 2: Results and comparison of SPA, Scree plot, Total variance, UCL, DBCEigen, RCEigen and Top-RA in the FERET database.

	SPA-u	SPA-p	Scree	Total var.	UCL	DBCEigen	RCEigen	Top-RA
$k = 1$	2	2	3	5	14	7	7	3
$k = 2$	2	2	3	6	15	7	13	2
$k = 3$	2	2	3	6	16	7	11	2
$k = 4$	2	2	3	6	16	7	16	6
$k = 5$	2	3	3	6	16	7	11	13
stability	0	0.4472	0	0.4472	0.8944	0	3.2863	4.6583
ATA	54.5%	54.3%	54.5%	53.2%	50.4%	52.0%	50.4%	56.2%
overall score	27.3	25.7	18.2	9.2	4.5	7.5	4.6	18.1

other methods. However, among them, SPA with upper-edge comparison selects the smallest number of PCs. Total variance, UCL, DBCEigen and RCEigen select more PCs with lower average testing accuracy. Regarding recognition accuracy and storage requirements, SPA with upper-edge comparison is the best.

4.3 Experiments on the Extended Yale B Database

The extended Yale Face B Database proposed by Georghiades et al. (2001) contains 16128 images of 28 human subjects under 9 poses and 64 illumination conditions. The image size is 480×640 pixels, much larger than that of images in the ORL and FERET datasets. A subset of the database is used for our experiment. We select 20 people and 20 images for each person. In contrast to the ORL and FERET databases, the pose and illumination are varied. The images are all the low quality, as a large part of an image is a heterogeneous background in the Yale Face Database B. Five sample images of one person from the Extended Yale B database are shown in Figure 7.

This Yale B dataset is used to evaluate the performance of SPA under conditions where the images are large-size and low-quality. We split the data with 10, 20, 30, 40, and 50 percent per

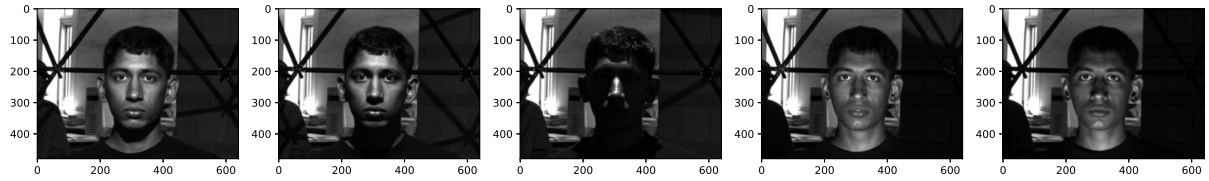


Figure 7: Five sample images of one person in the Extended Yale B database.

class (per person) for training and the rest data for testing. Thus, the total training samples are 40, 80, 120, 160, and 200, respectively, and the corresponding testing samples are 360, 320, 280, 240, and 200, respectively. The image covariance matrix G is calculated on the training samples for each case. To ensure the consistency of notation, we use $k = 1, 2, 3, 4, 5$ to denote these five cases.

First, the results of SPA are presented in Figure 8. The number of trials is $T = 20$. The five plots on the left panel show the first 60 largest eigenvalues of G of training samples (original data) and their signflipped analogs (permuted data) for cases $k = 1, 2, 3, 4, 5$, respectively. Although the magnitude of the eigenvalues is large for the first few, it still converges to zero after that. SPA algorithm with upper-edge comparison selects the first 8 largest eigenvalues for four cases $k = 1, 2, 3, 4$, and the first 9 largest eigenvalues in the case $k = 5$. SPA with pairwise comparison selects few more eigenvalues, and they are 10, 9, 9, 9, and 11 for five cases, respectively.

To illustrate the Top-RA method, the five plots on the right panel of Figure 8 present the classification accuracy of 2DPCA on the test data against different numbers of PCs kept in the training model. As the number of PCs kept increases, the testing classification accuracy increases and gradually levels off. For $k = 1, 2, 3, 4, 5$, the optimal numbers of PCs with top testing classification accuracy are 25, 38, 24, 24, and 24 for five cases, respectively. The vertical lines locate the accuracy of 8 PCs selected by SPA with upper-edge comparison. After 8 PCs, including more eigenvalues only slightly increases accuracy but wastes more time and storage memory.

The scree plot is just the line of the original data in each plot on the left panel of Figure 8. It is not easy to identify the elbow of the scree plots, and 5-9 are all possible choices. By setting the threshold as 0.8, the total variance chooses 12, 11, 11, 11, and 11 numbers of PCs for five cases, respectively. By setting the threshold as 0.1 and the significance level as $\alpha = 0.05$, the UCL selects 23, 22, 22, 22, and 23 numbers of PCs for five cases, respectively. The differences and ratios of consecutive eigenvalues are plotted in Figure 9, respectively, and they are more fluctuating compared to that in Figure 3 and Figure 6. It is hard to find optimal thresholds for DBCEigen and RCEigen subjectly. The thresholds for DBCEigen and RCEigen are subjectively set as 10^6 and 1.25, respectively. Both of them select 18 PCs for all cases.

For comparison, Table 3 summarizes the above results of these methods and the metrics of their performances in terms of stability, ATA and over score. In terms of stability, DBCEigen and RECigen are the best, but the selected number 18 is larger than that of SPA, scree plot, and total variance, which have very low variance. Regarding ATA, all methods are close to the optimal level in the Top-RA. However, UCL, DBCEigen, RCEigen and Top-RA have more number of PCs in the training model. This phenomenon is consistent with that observed on the right panel of Figure 8, that is, the testing accuracy increases slowly as more PCs are included. Overall, SPA with upper-edge comparison is the best one, considering both recognition accuracy and storage requirements.

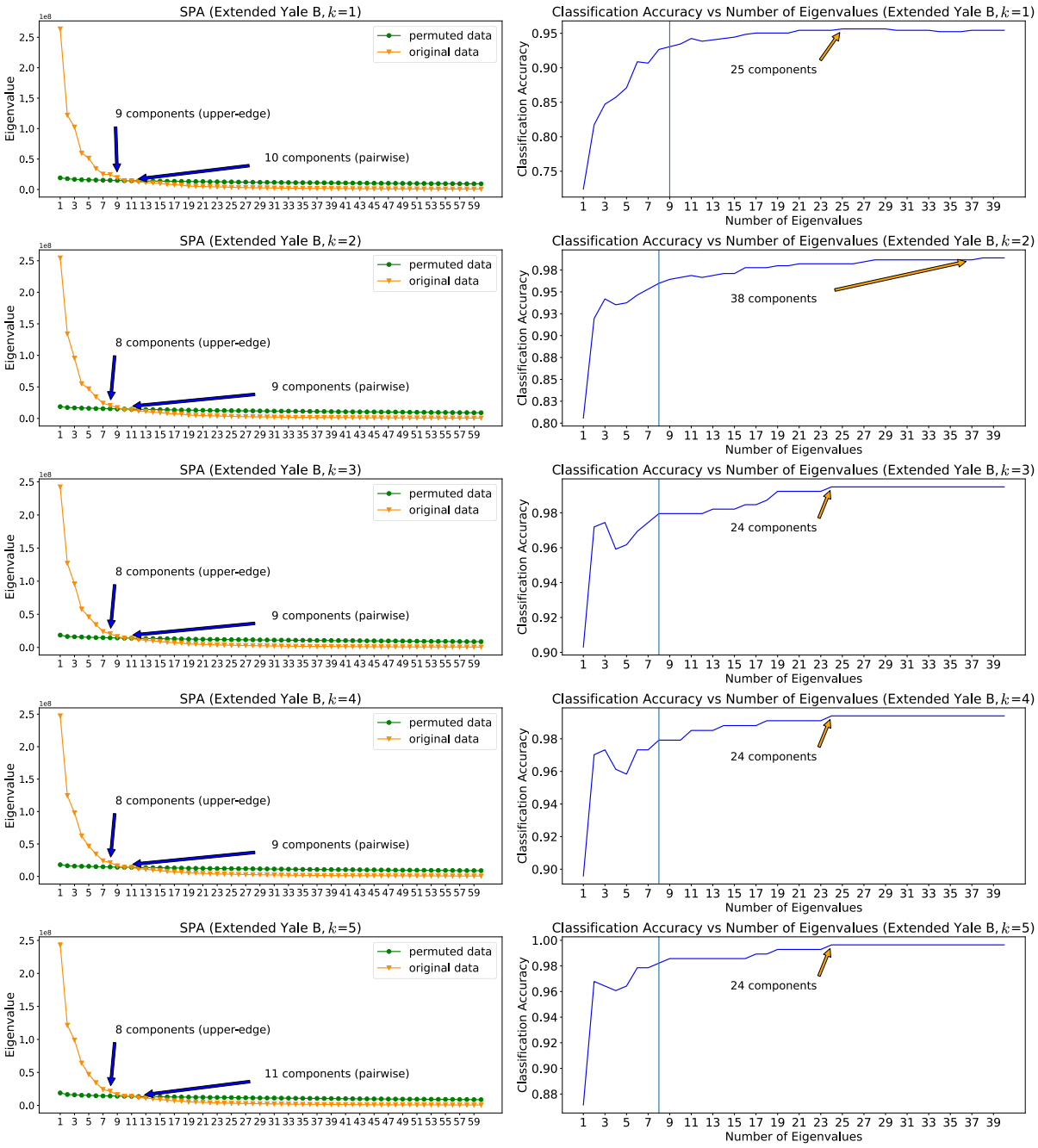


Figure 8: For the Yale B database. Left panel: number of PCs selected by SPA; Right panel: classification accuracy on test data against different numbers of PCs.

Last, the consuming times of running the SPA algorithm (including SPA with upper-edge comparison and SPA with pairwise comparison) to obtain the results on the left panel of Figures 2, 5, and 8 for the three databases, respectively, are summarized in Table 4. In comparison, the computing times of Top-RA have provided. Therefore, SPA is very efficient.

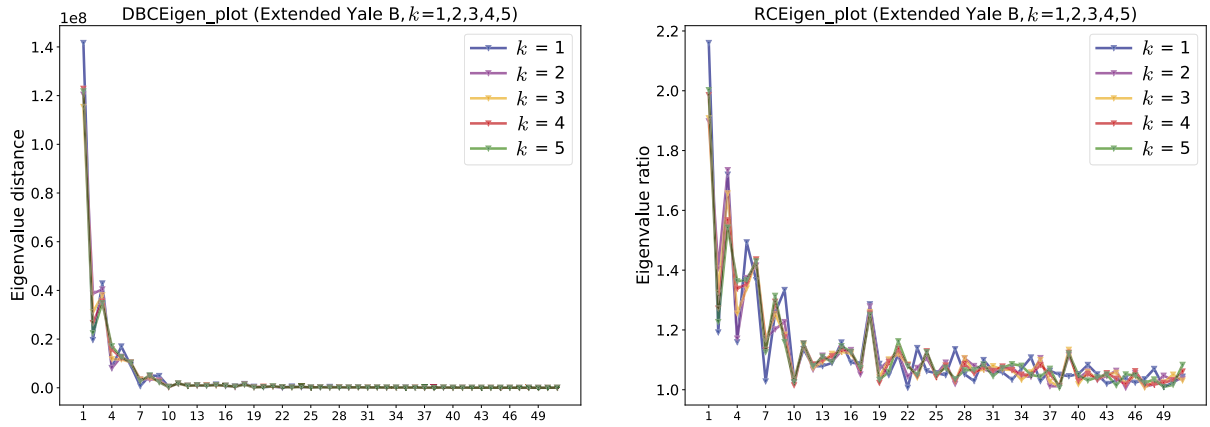


Figure 9: For the Yale B database. Left: differences between consecutive eigenvalues; Right: ratios of consecutive eigenvalues.

Table 3: Results and comparison of SPA, Scree plot, Total variance, UCL, DBCEigen, RCEigen and Top-RA in the Extended Yale B database. (8 PCs were used to calculate the three metrics for the scree plot, but the number of PCs may vary from person to person.)

	SPA-u	SPA-p	Scree	Total var.	UCL	DBCEigen	RCEigen	Top-RA
$k = 1$	9	10	5–9	12	23	18	18	25
$k = 2$	8	9	5–9	11	22	18	18	38
$k = 3$	8	9	5–9	11	22	18	18	24
$k = 4$	8	11	5–9	11	22	18	18	24
$k = 5$	8	10	5–9	11	23	18	18	24
stability	0.4472	0.8367	0*	0.4472	0.5477	0	0	6.1644
ATA	96.6%	96.9%	96.8%	97.2%	98.3%	98.0%	98.0%	98.6%
overall score	11.8	10.2	10.8	8.7	4.4	5.4	5.4	3.8

Table 4: Time consuming of SPA and Top-RA. (CPU: AMD Ryzen 7 6800H with Radeon Graphics, 8 cores, 16 threads; RAM: 16 GB; Operating system: Windows 11, 64-bit.)

	SPA			Top-RA		
	ORL	FERET	Yale B	ORL	FERET	Yale B
T	100	50	20	–	–	–
Size of image	92×112	80×80	480×640	92×112	80×80	480×640
k	1–5	1–5	1–5	1–5	1–5	1–5
Time (s)	18.2	12.7	125	281	3893	15814

5 Conclusion

In this paper, a novel yet simple algorithm named SPA is proposed to optimize the number of principal components of 2DPCA in the training set. As analyzed in Section 2, some traditional

methods are not suitable for image data or are subjective, but the existing method designed to choose the number of PCs in 2DPCA, i.e., depending on the top accuracy of the testing set, is not a stable and reliable way to determine the number of PCs in the training set. To solve this issue, we introduce a simple but effective method to optimize the number of PCs based on sign-flip parallel analysis. Specifically, some sign-flip permutation trials are conducted to characterize the noise level; therefore, the signals can be identified. Extensive experiments are conducted to evaluate the effectiveness of SPA. As demonstrated, SPA not only determines the rank of the training model with a significant statistical level, but is also more stable and the best regarding testing accuracy and storage requirements compared to the state-of-the-art. Therefore, this proposed method generalizes the application of 2DPCA in image representation and pattern recognition. It is also worth noting that the proposed method can be applied to any situation where the number of PCs is needed to determine training models, such as eigenfaces (Turk and Pentland, 1991), eigenhill (Yilmaz and Gokmen, 2000) and other eigenvector-based methods.

In SPA, the sign-flip permutation trial is applied to characterize the noise level; the number of trials significantly influences the performance of the proposed method. It will be valuable to investigate theoretical properties of SPA, including constructing the consistency of the estimated number of PCs, and determining the best number of trials.

Supplementary Material

The supplementary material contains a zipped folder, which contains codes and three data sets for reproducing all results. Please go to https://figshare.com/s/824176b60a12b8ee0535.

Acknowledgement

We are grateful to David Zhang for helpful comments.

Funding

Zhaoyuan Li's research is partially supported by National Natural Science Foundation of China (No. 11901492) and Shenzhen Science and Technology Program (ZDSYS 20211021111415025).

References

Ahn SC, Horenstein AR (2013). Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3): 1203–1227. https://doi.org/10.3982/ECTA8968

Bai Z, Silverstein JW (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer, New York.

Buja A, Eyuboglu N (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4): 509–540. https://doi.org/10.1207/s15327906mbr2704_2

Cattell RB, Vogelmann S (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, 12(3): 289–325. https://doi.org/10.1207/s15327906mbr1203_2

Dhahri H, Al Maghayreh E, Mahmood A, Elkilani W, Faisal Nagi M (2019). Automated breast cancer diagnosis based on machine learning algorithms. *Journal of Healthcare Engineering*, 2019(1): 4253641.

Ejaz MS, Islam MR, Sifatullah M, Sarker A (2019). Implementation of principal component analysis on masked and non-masked face recognition. In: *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 1–5. IEEE.

Georghiades AS, Belhumeur PN, Kriegman DJ (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6): 643–660. https://doi.org/10.1109/34.927464

Gumaei A, Hassan MM, Hassan MR, Alelaiwi A, Fortino G (2019). A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. *IEEE Access*, 7: 36266–36273. https://doi.org/10.1109/ACCESS.2019.2904145

Hair JF Jr, Anderson RE, Tatham RL (1986). *Multivariate Data Analysis with Readings*. Macmillan Publishing Co., Inc.

Hong D, Sheng Y, Dobriban E (2020). Selecting the number of components in PCA via random signflips. *arXiv preprint arXiv:2012.02985*.

Horn JL (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30: 179–185. https://doi.org/10.1007/BF02289447

Lam C, Yao Q (2012). Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 694–726.

Onatski A (2010). Determining the number of factors from empirical distribution of eigenvalues. *Review of Economics and Statistics*, 92(4): 1004–1016. https://doi.org/10.1162/REST_a_00043

Owen AB, Wang J (2016). Bi-cross-validation for factor analysis. *Statistical Science*, 31(1): 119–139. https://doi.org/10.1214/15-STS539

Steven Eyobu O, Han DS (2018). Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network. *Sensors*, 18(9): 2892. https://doi.org/10.3390/s18092892

Turk MA, Pentland AP (1991). Face recognition using eigenfaces. In: *Proceedings of 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 586–587. IEEE Computer Society.

Uddin MP, Mamun MA, Hossain MA (2021). PCA-based feature reduction for hyperspectral remote sensing image classification. *IETE Technical Review*, 38(4): 377–396. https://doi.org/10.1080/02564602.2020.1740615

Wan S, Xia Y, Qi L, Yang YH, Atiquzzaman M (2020). Automated colorization of a grayscale image with seed points propagation. *IEEE Transactions on Multimedia*, 22(7): 1756–1768. https://doi.org/10.1109/TMM.2020.2976573

Wang H (2012). Factor profiled sure independence screening. *Biometrika*, 99(1): 15–28. https://doi.org/10.1093/biomet/asr074

Wang P, Li Z, Wei Z, Wu T, Luo C, Jiang W, et al. (2024). Space-time-coding digital metasurface element design based on state recognition and mapping methods with CNN-LSTM-DNN. *IEEE Transactions on Antennas and Propagation*, 72(6): 4962–4975. https://doi.org/10.1109/TAP.2024.3349778

Wang Q, Gao Q, Gao X, Nie F (2017). Optimal mean two-dimensional principal component analysis with F-norm minimization. *Pattern Recognition*, 68: 286–294. https://doi.org/10.1016/j.patcog.2017.03.026

Yang J, Zhang D, Frangi AF, Jy Y (2004). 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. https://doi.org/10.1109/TPAMI.2004.1261097

Yang W, Wang S, Hu J, Tao X, Li Y (2024). Feature extraction and learning approaches for cancellable biometrics: a survey. *CAAI Transactions on Intelligence Technology*, 9(1): 4–25. https://doi.org/10.1049/cit2.12283

Yilmaz A, Gokmen M (2000). Eigenhill vs. eigenface and eigenedge. In: *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, 827–830. IEEE.

Zabalza J, Ren J, Yang M, Zhang Y, Wang J, Marshall S, et al. (2014). Novel folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 93: 112–122. https://doi.org/10.1016/j.isprsjprs.2014.04.006

Zeng X, Wang X, Xie Y (2024). Multiple pseudo-siamese network with supervised contrast learning for medical multi-modal retrieval. *ACM Transactions on Multimedia Computing Communications and Applications*, 20(5): 1–23.