Constrained Projection Approximation Algorithms for Principal Component Analysis

Seungjin Choi[§], Jong-Hoon Ahn[†], Andrzej Cichocki[‡]
[§] *Department of Computer Science, Pohang University of Science and Technology, San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea. e-mail: seungjin@postech.ac.kr*
[†] *Department of Physics, Pohang University of Science and Technology, San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Korea. e-mail: jonghun@postech.ac.kr*
[‡] *Advanced Brain Signal Processing Lab, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan. e-mail: cia@brain.riken.jp*

May 30, 2006

Abstract. In this paper we introduce a new error measure, *integrated reconstruction error* (IRE) and show that the minimization of IRE leads to principal eigenvectors (without rotational ambiguity) of the data covariance matrix. Then we present iterative algorithms for the IRE minimization, where we use the projection approximation. The proposed algorithm is referred to as COnstrained Projection Approximation (COPA) algorithm and its limiting case is called COPAL. Numerical experiments demonstrate that these algorithms successfully find exact principal eigenvectors of the data covariance matrix.

Keywords: Natural power iteration, principal component analysis, projection approximation, reconstruction error, subspace analysis.

1. Introduction

Principal component analysis (PCA) or principal subspace analysis (PSA) is a fundamental multivariate data analysis method which is encountered into a variety of areas in neural networks, signal processing, and machine learning (Jolliffe, 2002). A variety of adaptive (on-line) algorithms for PCA or PSA can be found in literature (Oja, 1989; Baldi and Hornik, 1989; Sanger, 1989; Brockett, 1991; Cichocki and Unbehauen, 1992; Xu, 1993). See also (Diamantaras and Kung, 1996) and references therein. Most of these algorithms are gradient-based learning algorithms, hence the convergence is slow.

The power iteration is a classical method for estimating the largest eigenvector of a symmetric matrix. The subspace iteration is a direct extension of the power iteration, computing subspace spanned by principal eigenvectors of a symmetric matrix. The natural power method is an exemplary instance of the subspace iteration, where the invariant subspace spanned by the n largest eigenvectors of the data covariance matrix, is determined (Hua et al., 1999). The natural power iteration

provides a general framework for several well-known subspace algorithms, including Oja's subspace rule (Oja, 1989), PAST (Yang, 1995), and OPAST (Abed-Meraim et al., 2000).

A common derivation of PSA, is terms of a linear (orthogonal) projection $\boldsymbol{W} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_n] \in \mathbb{R}^{m\times n}$ such that given a centered data matrix $\boldsymbol{X} = [\boldsymbol{x}(1) \cdots \boldsymbol{x}(N)] \in \mathbb{R}^{m\times N}$, the reconstruction error $\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{W}^\top \boldsymbol{X}\|_F^2$ is minimized, where $\|\cdot\|_F$ denotes the Frobenius norm (Euclidean norm). It is known that the reconstruction error is blind to an arbitrary rotation of the representation space. The minimization of the reconstruction error leads to $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{Q}$ where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is an arbitrary orthogonal matrix and the eigendecomposition of the covariance matrix $\boldsymbol{C} = \boldsymbol{X}\boldsymbol{X}^\top$ is given by

$$\boldsymbol{C} = [\ \boldsymbol{U}_1\ \ \boldsymbol{U}_2\ ] \begin{bmatrix} \boldsymbol{\Lambda}_1 & 0 \\ 0 & \boldsymbol{\Lambda}_2 \end{bmatrix} [\ \boldsymbol{U}_1\ \ \boldsymbol{U}_2\ ]^\top, \tag{1}$$

where $\boldsymbol{U}_1 \in \mathbb{R}^{m\times n}$ contains n largest eigenvectors, $\boldsymbol{U}_2 \in \mathbb{R}^{m\times(m-n)}$ consists of the rest of eigenvectors, and associated eigenvalues are in $\boldsymbol{\Lambda}_1,\ \boldsymbol{\Lambda}_2$ with $\lambda_1 > \lambda_2 > \cdots > \lambda_m$.

Probabilistic model-based method for PCA was developed, where the linear generative model was considered and expectation maximization (EM) optimization was used to derive iterative PCA algorithms, including probabilistic PCA (PPCA) (Tipping and Bishop, 1999) and EM-PCA (Roweis, 1998). These algorithms are batch algorithms that find principal subspace. Hence, further post-processing is required to determine exact principal eigenvectors of the data covariance matrix, without rotational ambiguity. The natural power iteration is also a PSA-type algorithm, unless the deflation method is used.

In this paper we present iterative algorithms which determine the principal eigenvectors of the data covariance matrix in a parallel fashion (in contrast to the deflation method). To this end, we first introduce the *integrated reconstruction error* (IRE) and show that its minimization leads to exact principal eigenvectors (without rotational ambiguity). Proposed iterative algorithms emerge from the minimization of the IRE and are referred to as COnstrained Projection Approximation (COPA) algorithm and COPAL (the limiting case of COPA). These algorithms are the recognition model counterpart of the constrained EM algorithm in (Ahn and Oh, 2003; Ahn et al., 2004) where principal directions are estimated through alternating two steps (E and M steps) in the context of the linear coupled generative model. In contrast, our algorithms COPA and COPAL, need not go through two steps, which is a major advantage over EM type algorithms.

2. Integrated Reconstruction Error

It was shown in (Yang, 1995) that the reconstruction error $\mathcal{J}_{RE} = \|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{X}\|_F^2$ attains the global minimum if and only if $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{Q}$. Now we introduce the IRE that is summarized below.

DEFINITION 1 (IRE). *The integrated reconstruction error, $\mathcal{J}_{IRE}$, is defined as a linear combination of n partial reconstruction errors (PRE), $\mathcal{J}_i = \left\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{E}_i\boldsymbol{W}^\top\boldsymbol{X}\right\|_F^2$, i.e.,*

$$\begin{aligned}\mathcal{J}_{IRE}(\boldsymbol{W}) &= \sum_{i=1}^{n}\alpha_i\mathcal{J}_i \\ &= \sum_{i=1}^{n}\alpha_i\left\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{E}_i\boldsymbol{W}^\top\boldsymbol{X}\right\|^2, \end{aligned} \tag{2}$$

where coefficients α_i are positive real numbers and $\boldsymbol{E}_i \in \mathbb{R}^{n\times n}$ is a diagonal matrix, defined by

$$[\boldsymbol{E}_i]_{jj} = \begin{cases} 1 & \text{for } j = 1,\ldots,i, \\ 0 & \text{for } j = i+1,\ldots,n. \end{cases}$$

THEOREM 1 (Main Theorem). *The IRE is minimized if and only if $\boldsymbol{W} = \boldsymbol{U}_1$.*

Proof. See Appendix A.

Remarks:

- The last term in IRE, $\mathcal{J}_n$, is the standard reconstruction error. It was shown in (Yang, 1995) that $\boldsymbol{W}$ is a stationary point of $\mathcal{J}_n$ if and only if $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{Q}$ (hence $\boldsymbol{W}^\top\boldsymbol{W} = \boldsymbol{I}$ is satisfied). All stationary points of $\mathcal{J}_n$ are saddle points, except when $\boldsymbol{U}_1$ contains the n dominant eigenvectors of $\boldsymbol{C}$. In that case, $\mathcal{J}_n$ attains the global minimum.
- The standard reconstruction error $\mathcal{J}_n$ is invariant to an orthogonal transform $\boldsymbol{Q}$ because $\boldsymbol{W}\boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{W}^\top$. In contrast, the IRE is not invariant under an orthogonal transform, since $\boldsymbol{Q}\boldsymbol{E}_i\boldsymbol{Q}^\top \neq \boldsymbol{E}_i$. This provides an intuitive idea why the IRE minimization leads to the principal eigenvectors of the data covariance matrix, without rotational ambiguity.
- PREs $\mathcal{J}_i$ are of the form

$$\mathcal{J}_i = \left\|\boldsymbol{X} - \left(\boldsymbol{w}_1\boldsymbol{w}_1^\top + \cdots + \boldsymbol{w}_i\boldsymbol{w}_i^\top\right)\boldsymbol{X}\right\|_F^2,$$

where $\boldsymbol{w}_i$ represents the ith column vector of $\boldsymbol{W}$. The PRE $i+1$, $\mathcal{J}_{i+1}$, represents the reconstruction error for $(i+1)$-dimensional principal subspace which completely includes i-dimensional principal subspace. Therefore the minimization of $\mathcal{J}_{IRE}$ implies that each PRE $\mathcal{J}_i$ for $i = 1, \ldots, n$, is minimized. The graphical representation is shown in Fig 1, where a coupled linear recognition model is described, with a link of the IRE minimization.

– Minimizing each $\mathcal{J}_i$ is reminiscent of the deflation method where the eigenvectors of $\boldsymbol{C}$ are extracted one by one. Thus, it is expected that the minimization of IRE leads to principal eigenvectors of $\boldsymbol{C}$. However, a major difference between the deflation method and our method is that the former extracts principal components one by one and the latter find principal components simultaneously.

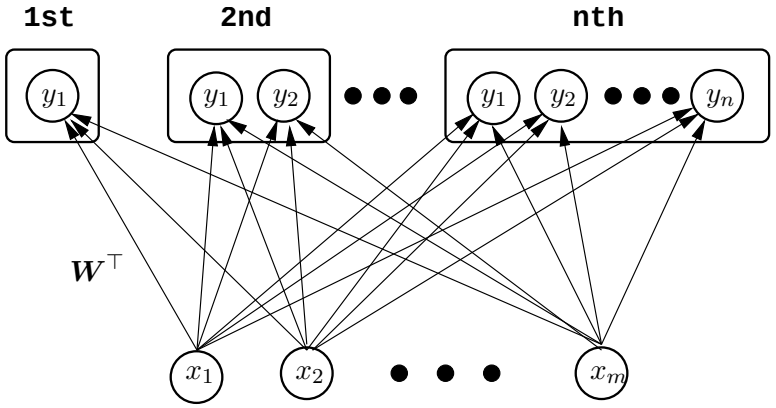


Figure 1. A coupled linear recognition model is shown, where it consists of n sub-models coupled through sharing the weights arriving at the same y_i for each sub-model. The ith sub-model is described by $y_j = \sum_{l=1}^{m} W_{lj} x_l = \boldsymbol{w}_j^\top \boldsymbol{x}$ for $j = 1, \ldots, i$. The reconstruction error for the ith model is given by $\mathcal{J}_i$. The weight matrix $\boldsymbol{W}$ is learned in such a way that the reconstruction errors, $\mathcal{J}_1, \ldots, \mathcal{J}_n$ are minimized.

3. Iterative Algorithms

The projection approximation (Yang, 1995) assumes that the difference between $\boldsymbol{W}_{(k+1)}^\top \boldsymbol{X}$ and $\boldsymbol{W}_{(k)}^\top \boldsymbol{X}$ is small, which leads us to consider the following objective function

$$\mathcal{J}_{IRE}(\boldsymbol{W}) = \frac{1}{2} \sum_{i=1}^{n} \alpha_i \left\| \boldsymbol{X} - \boldsymbol{W}_{(k+1)} \boldsymbol{E}_i \boldsymbol{Y}_{(k)} \right\|^2, \tag{3}$$

where $\boldsymbol{Y}_{(k)} = \boldsymbol{W}_{(k)}^\top \boldsymbol{X}$.

The gradient of (3) with respect to $\boldsymbol{W}_{(k+1)}$ is given by

$$\frac{\partial \mathcal{J}_{IRE}}{\partial \boldsymbol{W}_{(k+1)}} = -\boldsymbol{X}\boldsymbol{Y}_{(k)}^{\top}\boldsymbol{\Sigma} + \boldsymbol{W}_{(k+1)}\left[\left(\boldsymbol{Y}_{(k)}\boldsymbol{Y}_{(k)}^{\top}\right) \odot \boldsymbol{\Gamma}\right], \tag{4}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sum_{i=1}^{n}\alpha_i & 0 & 0 & \cdots & 0 \\ 0 & \sum_{i=2}^{n}\alpha_i & 0 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_n \end{bmatrix},$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \sum_{i=1}^{n}\alpha_i & \sum_{i=2}^{n}\alpha_i & \sum_{i=3}^{n}\alpha_i & \cdots & \alpha_n \\ \sum_{i=2}^{n}\alpha_i & \sum_{i=2}^{n}\alpha_i & \sum_{i=3}^{n}\alpha_i & \cdots & \alpha_n \\ \sum_{i=3}^{n}\alpha_i & \sum_{i=3}^{n}\alpha_i & \sum_{i=3}^{n}\alpha_i & \cdots & \alpha_n \\ \vdots & & & \ddots & \vdots \\ \alpha_n & \alpha_n & \alpha_n & \cdots & \alpha_n \end{bmatrix},$$

and $\odot$ is the Hadamard product (element-wise product).

With these definitions, it follows from $\frac{\partial \mathcal{J}_{IRE}}{\partial \boldsymbol{W}_{(k+1)}} = 0$ that

$$\begin{aligned} \boldsymbol{W}_{(k+1)} &= \left[\boldsymbol{X}\boldsymbol{Y}_{(k)}^{\top}\right]\boldsymbol{\Sigma}\left[\left(\boldsymbol{Y}_{(k)}\boldsymbol{Y}_{(k)}^{\top}\right) \odot \boldsymbol{\Gamma}\right]^{-1} \\ &= \left[\boldsymbol{X}\boldsymbol{Y}_{(k)}^{\top}\right]\left[\left(\boldsymbol{Y}_{(k)}\boldsymbol{Y}_{(k)}^{\top}\right) \odot \left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\right)\right]^{-1} \\ &= \boldsymbol{X}\boldsymbol{Y}_{(k)}^{\top}\left[\mathsf{U}\left(\boldsymbol{Y}_{(k)}\boldsymbol{Y}_{(k)}^{\top}\right)\right]^{-1}, \end{aligned} \tag{5}$$

where $\mathsf{U}(\boldsymbol{Y})$ is an element-wise operator, whose arguments Y_{ij} are transformed by

$$\mathsf{U}(Y_{ij}) = \begin{cases} Y_{ij}\dfrac{\sum_{l=i}^{n}\alpha_l}{\sum_{l=j}^{n}\alpha_l} & \text{if } i > j \\ Y_{ij} & \text{if } i \le j \end{cases}. \tag{6}$$

The operator $\mathsf{U}(\boldsymbol{Y})$ results from the structure of $\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}$ given by

$$\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ \frac{\sum_{i=2}^{n}\alpha_i}{\sum_{i=1}^{n}\alpha_i} & 1 & 1 & \cdots & 1 \\ \frac{\sum_{i=3}^{n}\alpha_i}{\sum_{i=1}^{n}\alpha_i} & \frac{\sum_{i=3}^{n}\alpha_i}{\sum_{i=2}^{n}\alpha_i} & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ \frac{\alpha_n}{\sum_{i=1}^{n}\alpha_i} & \frac{\alpha_n}{\sum_{i=2}^{n}\alpha_i} & \frac{\alpha_n}{\sum_{i=3}^{n}\alpha_i} & \cdots & 1 \end{bmatrix}.$$

Replacing $\boldsymbol{Y}_{(k)}$ by $\boldsymbol{W}_{(k)}^{\top}\boldsymbol{X}$, leads to the updating rule for COPA:

$$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)}\left[\mathsf{U}\left(\boldsymbol{W}_{(k)}^{\top}\boldsymbol{C}\boldsymbol{W}_{(k)}\right)\right]^{-1}. \tag{7}$$

In the limit of $\frac{\alpha_{i+1}}{\alpha_i} \to 0$ for $i = 1, \ldots, n-1$, $\mathsf{U}(\cdot)$ becomes the conventional upper-triangularization operator U_T which is given by

$$\mathsf{U}_T(Y_{ij}) = \begin{cases} 0 & \text{if } i > j, \\ Y_{ij} & \text{if } i \le j. \end{cases} \tag{8}$$

This leads to the COPAL algorithm

$$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)}\left[\mathsf{U}_T\left(\boldsymbol{W}_{(k)}^{\top}\boldsymbol{C}\boldsymbol{W}_{(k)}\right)\right]^{-1}. \tag{9}$$

Algorithms are summarized in Table I, where the constrained natural power iteration is a variation of the natural power iteration (Hua et al., 1999), while incorporating with the upper-triangularization operator U_T. The validity of the COPAL algorithm is justified by the following theorem where the fixed point of (9) is shown to correspond to the eigenvector matrix $\boldsymbol{U}_1$.

THEOREM 2. *The fixed point $\boldsymbol{W}$ of the COPAL (9) satisfies $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{\Upsilon}$ (after each column vector of $\boldsymbol{W}$ is normalized), where $\boldsymbol{\Upsilon}$ is a diagonal matrix with its diagonal entries being 1 or -1, provided that the nth and $(n+1)$th eigenvalues of $\boldsymbol{C}$ are distinct and the initial weight matrix $\boldsymbol{W}_{(0)}$ meets a mild condition, saying that there exists a nonsingular matrix $\boldsymbol{L} \in \mathbb{R}^{(m-n)\times n}$ such that $\boldsymbol{U}_2^{\top}\boldsymbol{W}_{(0)} = \boldsymbol{L}\boldsymbol{U}_1^{\top}\boldsymbol{W}_{(0)}$ for a randomly chosen $\boldsymbol{W}_{(0)}$.*

Table I. The outline of updating rules and the characteristics of algorithms, is summarized, where PAST and NP (natural power) are given as their batch version and CNP stands for the constrained natural power.

Algorithm	Updating rule	Type
PAST (Yang, 1995)	$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)} \left[\boldsymbol{W}_{(k)}^\top \boldsymbol{C}\boldsymbol{W}_{(k)}\right]^{-1}$	PSA
NP (Hua et al., 1999)	$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)} \left[\boldsymbol{W}_{(k)}^\top \boldsymbol{C}^2\boldsymbol{W}_{(k)}\right]^{-\frac{1}{2}}$	PSA
CNP (Choi, 2005)	$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)} \left[\mathsf{U}_T\left(\boldsymbol{W}_{(k)}^\top \boldsymbol{C}^2\boldsymbol{W}_{(k)}\right)\right]^{-\frac{1}{2}}$	PCA
COPA	$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)} \left[\mathsf{U}\left(\boldsymbol{W}_{(k)}^\top \boldsymbol{C}\boldsymbol{W}_{(k)}\right)\right]^{-1}$	PCA
COPAL	$\boldsymbol{W}_{(k+1)} = \boldsymbol{C}\boldsymbol{W}_{(k)} \left[\mathsf{U}_T\left(\boldsymbol{W}_{(k)}^\top \boldsymbol{C}\boldsymbol{W}_{(k)}\right)\right]^{-1}$	PCA

Proof. See Appendix B.

4. Numerical Experiments

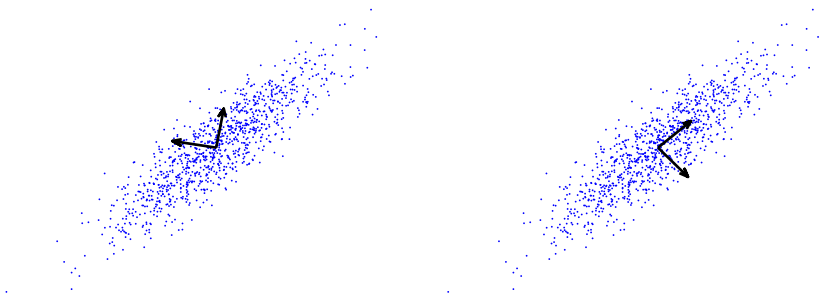
Numerical examples are provided, in order to verify that the weight matrix $\boldsymbol{W}$ in COPA as well as COPAL converges to the true eigenvectors of the data covariance matrix $\boldsymbol{C}$.

4.1. Example 1

The first experiment was carried out with 2-dimensional vector sequences of length 1000. Fig. 2 shows the data scatter plots and principal directions computed by the PAST algorithm and by our algorithms (COPA and COPAL). One can see that principal directions estimated by the PAST algorithm are rotated eigenvectors of the data covariance matrix (i.e., principal subspace). On the other hand, COPA or COPAL finds exact principal directions (see Fig. 2 (b)).

4.2. Example 2

In this example, we show different convergence behavior of the COPA algorithm, depending on the choice of α_i. Regardless of the values of α_i, the minimum of the IRE stays the same. However, the convergence behavior of the COPA algorithm is different, especially according to the ratio $\frac{\alpha_{i+1}}{\alpha_i}$ for $i = 1, \ldots, n-1$ (see Fig. 3). In this example, 5-dimensional Gaussian random vectors with 1000 samples, were linearly transformed to generate the 10-dimensional data matrix $\boldsymbol{X} \in \mathbb{R}^{10\times 1000}$. Fig 3 shows the convergence behavior of the COPA algorithm with different choice of α_i, as well as the COPAL algorithm. What was found here that



(a) (b)

Figure 2. Principal directions computed by: (a) PAST algorithm (or the natural power); (b) COPA (or COPAL). The PAST (or NP) algorithm finds rotated principal directions, whereas our algorithms (COPA and COPAL) estimate exact principal directions of the two-dimensional data.

the convergence of the COPA becomes faster, as the ratio, $\frac{\alpha_{i+1}}{\alpha_i}$ for $i = 1, \ldots, n-1$ decreases.

4.3. Example 3

This example involves the useful behavior of our algorithms for high-dimensional data, showing that even for the case of high-dimensional data, our algorithms successfully estimate exact first few principal directions of data. To this end, we generated 5000 5-dimensional Gaussian vectors (with zero mean and unit variance) and applied a linear transform to construct the data matrix $\boldsymbol{X} \in \mathbb{R}^{1000\times 5000}$. The rank of the covariance matrix $\boldsymbol{C}$ is 5. COPA and COPAL algorithms in (7) and (9) were applied to find 3 principal eigenvectors from this data matrix. For the case of COPA, we used $\alpha_1 = 1,\ \alpha_2 = 0.1,\ \alpha_3 = 0.01$. Results are shown in Fig. 4.

4.4. Example 4

As a real-world data example, we applied the COPAL algorithm to USPS handwritten digit data, in order to determine eigen-digits (see Fig. 5. Each image is the size of 16×16, which is converted to a 256-dimensional vector. First 100 principal components were estimated by the COPAL algorithm as well as SVD and the batch version of PASTd (PAST with deflation). Although the deflation method determines eigenvectors without rotational ambiguity, however, error accumulation is propagated as n increases.

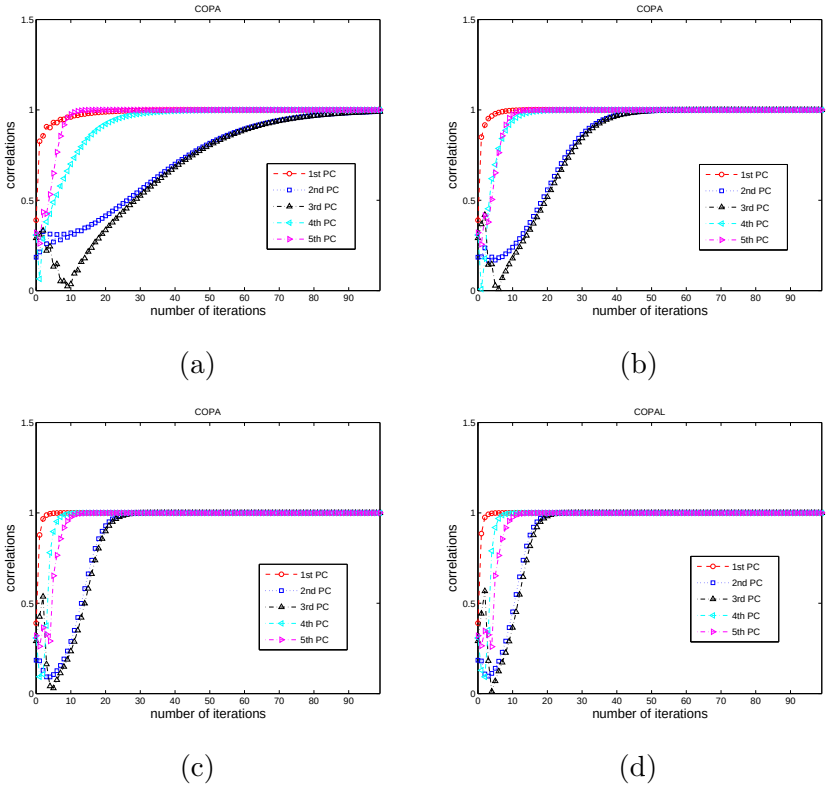


Figure 3. Evolution of weight vectors is shown in terms of the absolute value of the inner produce between a weight vector and a true eigenvector (computed by SVD): (a) COPA with $\frac{\alpha_{i+1}}{\alpha_i} = 1$ and $\alpha_1 = 1$; (b) COPA with $\frac{\alpha_{i+1}}{\alpha_i} = 0.5$ and $\alpha_1 = 1$; (c) COPA with $\frac{\alpha_{i+1}}{\alpha_i} = 0.1$ and $\alpha_1 = 1$; (d) COPAL.

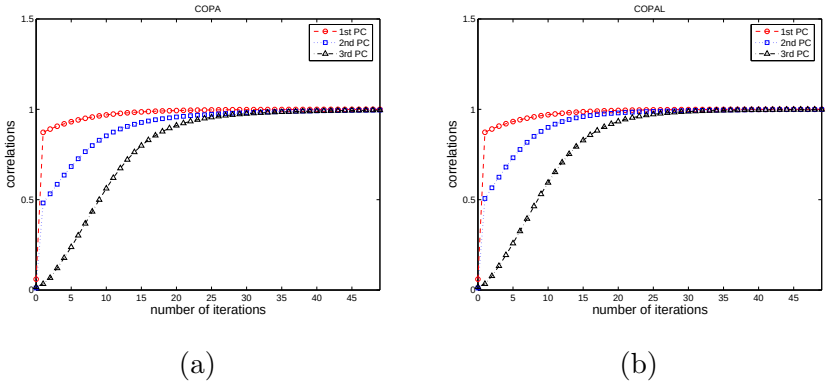


Figure 4. Evolution of weight vectors: (a) COPA; (b) COPAL. Correlations represent the absolute value of the inner product between a weight vector and a true eigenvector (computed by SVD).

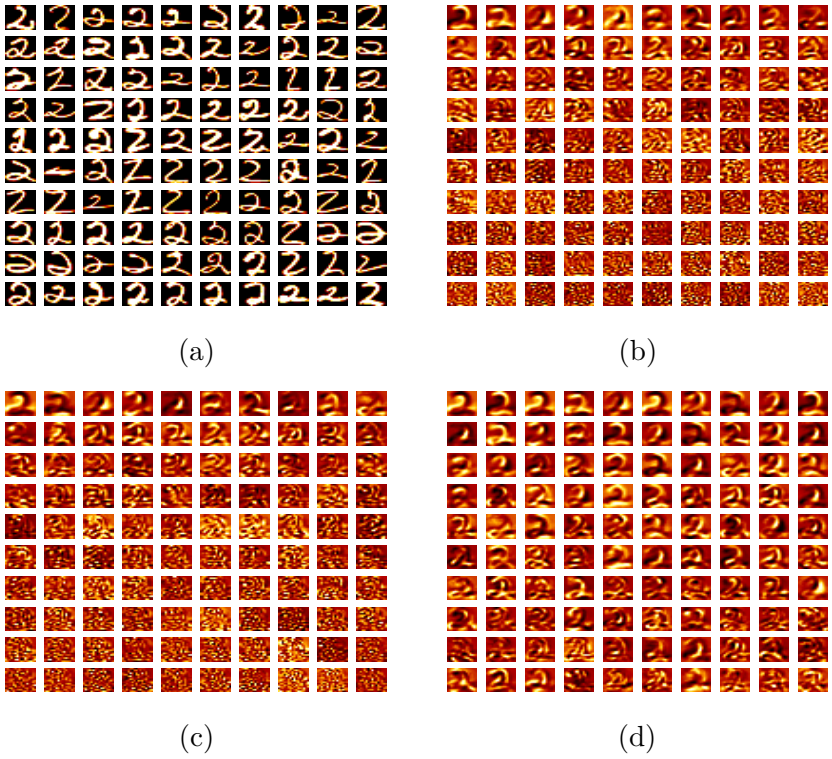


Figure 5. USPS hand-written digit data, '2', is shown in (a). The rest are corresponding principal components estimated by: (b) SVD; (c) COPAL; (d) PASTd (PAST with deflation). The eigen-digits estimated by COPAL is exactly same as ones found by SVD. On the other hand, first 10-20 eigen-digits computed by the deflation methtod are same as true eigen-digits, but eigen-digits are deteriorated as n increases.

5. Conclusions

We have presented two iterative algorithms, COPA and COPAL, which determine principal eigenvectors of the data covariance matrix. In contrast to PPCA, EM-PCA, PAST, and NP, the algorithms COPA and COPAL could determine the eigenvectors without rotational ambiguity, since they were derived from the minimization of the integrated reconstruction error that was introduced in this paper. The COPAL algorithm emerged as a limiting case of COPA and its fixed point analysis was provided. The validity of two algorithms was demonstrated through several numerical examples where a few principal eigenvectors were required to be computed from very high-dimensional data. The useful behavior of COPA and COPAL was also shown, compared to

the deflation method where eigenvectors of the data covariance matrix are extracted one by one.

Acknowledgments

This work was supported by KISTEP International Cooperative Research Program, ITEP Brain Neuroinformatics Program, and Korea MIC under ITRC support program supervised by the IITA (IITA-2005-C1090-0501-0018).

Appendix A: Proof of Main Theorem

The sufficiency (if part) can be proved in a straightforward manner. The necessity (only if part) is proved in an induction-like manner. As mentioned in Sec. 2, the IRE is minimized if and only if each PRE $\mathcal{J}_i$ is minimized, since the IRE is a linear sum of PREs with positive coefficients $\{\alpha_i\}$ and i-dimensional subspace (determined by the minimization of $\mathcal{J}_i$) is completely included in $(i+1)$-dimensional subspace. Recall that true normalized eigenvectors of $\boldsymbol{C}$ are denoted by $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ with associated eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_n$.

We first consider $\mathcal{J}_1$ and show that its minimization implies $\frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} = \boldsymbol{u}_1$. It follows from $\frac{\partial \mathcal{J}_1}{\partial \boldsymbol{W}} = 0$ that we have

$$\boldsymbol{C}\boldsymbol{w}_1 = \boldsymbol{w}_1 \left(\boldsymbol{w}_1^\top \boldsymbol{C} \boldsymbol{w}_1 \right), \tag{10}$$

which implies $\boldsymbol{w}_1 = \boldsymbol{u}_{(k)}$, i.e., $\boldsymbol{w}_1$ is one of the normalized eigenvectors of $\boldsymbol{C}$. Then $\mathcal{J}_1$ can be written as

$$\begin{aligned} \mathcal{J}_1 &= \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{E}_1\boldsymbol{W}^\top\boldsymbol{X} \right\|^2 \\ &= \operatorname{tr}\left\{ \left(\boldsymbol{I} - \boldsymbol{w}_1\boldsymbol{w}_1^\top \right) \boldsymbol{C} \left(\boldsymbol{I} - \boldsymbol{w}_1\boldsymbol{w}_1^\top \right)^\top \right\} \\ &= \sum_{i \neq k} \lambda_i, \end{aligned} \tag{11}$$

where the 3rd equality directly comes from the spectral decomposition of $\boldsymbol{C}$, replacing $\boldsymbol{C}$ by $\sum_{i=1}^n \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$. Since $\lambda_1 > \lambda_2 > \cdots > \lambda_n$, the $\mathcal{J}_1$ is minimized when $k = 1$. Hence, $\boldsymbol{w}_1 = \boldsymbol{u}_1$.

Suppose that the minimization of $\sum_{j=1}^i \alpha_j \mathcal{J}_j$ leads to $\boldsymbol{w}_j = \boldsymbol{u}_j$ for $j = 1, \ldots, i$. Then we show that $\boldsymbol{w}_{i+1} = \boldsymbol{u}_{i+1}$ emerges from the minimization of $\mathcal{J}_{i+1}$. Solving $\frac{\partial \mathcal{J}_{i+1}}{\partial \boldsymbol{W}} = 0$ for $\boldsymbol{W}$, leads to

$$\boldsymbol{C}\boldsymbol{W}\boldsymbol{E}_{i+1} = \boldsymbol{W}\boldsymbol{E}_{i+1}\boldsymbol{W}^\top\boldsymbol{C}\boldsymbol{W}\boldsymbol{E}_{i+1}, \tag{12}$$

which can be re-written as

$$\boldsymbol{C}\left[\boldsymbol{W}_i \;\; \boldsymbol{w}_{i+1}\right] \;=\; \left[\boldsymbol{W}_i \;\; \boldsymbol{w}_{i+1}\right] \begin{bmatrix} \boldsymbol{W}_i^\top \boldsymbol{C}\boldsymbol{W}_i & \boldsymbol{W}_i^\top \boldsymbol{C}\boldsymbol{w}_{i+1} \\ \boldsymbol{w}_{i+1}^\top \boldsymbol{C}\boldsymbol{W}_i & \boldsymbol{w}_{i+1}^\top \boldsymbol{C}\boldsymbol{w}_{i+1} \end{bmatrix}, \quad (13)$$

where $\boldsymbol{W}_i = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_i]$. It follows from (13) that we have

$$\boldsymbol{C}\boldsymbol{W}_i \;=\; \boldsymbol{W}_i \left(\boldsymbol{W}_i^\top \boldsymbol{C}\boldsymbol{W}_i\right) + \boldsymbol{w}_{i+1}\boldsymbol{w}_{i+1}^\top \boldsymbol{C}\boldsymbol{W}_i, \quad (14)$$

$$\boldsymbol{C}\boldsymbol{w}_{i+1} \;=\; \boldsymbol{W}_i \left(\boldsymbol{W}_i^\top \boldsymbol{C}\boldsymbol{w}_{i+1}\right) + \boldsymbol{w}_{i+1}\boldsymbol{w}_{i+1}^\top \boldsymbol{C}\boldsymbol{w}_{i+1}. \quad (15)$$

Note that the stationary points of $\mathcal{J}_i$ satisfy

$$\boldsymbol{C}\boldsymbol{W}_i \;=\; \boldsymbol{W}_i \left(\boldsymbol{W}_i^\top \boldsymbol{C}\boldsymbol{W}_i\right).$$

Taking this relation into account in (14), leads to the orthogonality

$$\boldsymbol{w}_{i+1}^\top \boldsymbol{C}\boldsymbol{W}_i = 0. \quad (16)$$

Taking this orthogonality into account in (15), leads to

$$\boldsymbol{C}\boldsymbol{w}_{i+1} \;=\; \boldsymbol{w}_{i+1} \left(\boldsymbol{w}_{i+1}^\top \boldsymbol{C}\boldsymbol{w}_{i+1}\right), \quad (17)$$

which implies that $\boldsymbol{w}_{i+1}$ is one of eigenvectors, $\{\boldsymbol{u}_{i+1}, \ldots, \boldsymbol{u}_n\}$. Once again using the spectral decomposition of $\boldsymbol{C}$, the $\mathcal{J}_{i+1}$ can be written as

$$\mathcal{J}_{i+1} = \sum_{j=i+1, j\neq k}^{n} \lambda_j. \quad (18)$$

Thus, $\mathcal{J}_{i+1}$ is minimized when $k = i+1$, leading to $\boldsymbol{w}_{i+1} = \boldsymbol{u}_{i+1}$. This proves the main theorem. ■

Appendix B: Proof of Theorem 2

We define $\boldsymbol{\Phi}_{(k)} = \boldsymbol{U}_1^\top \boldsymbol{W}_{(k)}$ and $\boldsymbol{\Omega}_{(k)} = \boldsymbol{U}_2^\top \boldsymbol{W}_{(k)}$. With these definitions, pre-multiplying both sides of (9) by $[\boldsymbol{U}_1 \boldsymbol{U}_2]^\top$ leads to

$$\begin{bmatrix} \boldsymbol{\Phi}_{(k+1)} \\ \boldsymbol{\Omega}_{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_1 & 0 \\ 0 & \boldsymbol{\Lambda}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_{(k)} \\ \boldsymbol{\Omega}_{(k)} \end{bmatrix} \boldsymbol{Z}_{(k)}, \quad (19)$$

where

$$\boldsymbol{Z}_{(k)} = \left\{ \mathsf{U}_T \left[\boldsymbol{\Phi}_{(k)}^\top \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_{(k)} + \boldsymbol{\Omega}_{(k)}^\top \boldsymbol{\Lambda}_2 \boldsymbol{\Omega}_{(k)} \right] \right\}^{-1}. \quad (20)$$

As in the convergence proof of the natural power iteration in (Hua et al., 1999), one can show that $\boldsymbol{\Omega}_{(k)}$ goes to zero. Assume that $\boldsymbol{\Phi}_{(0)} \in \mathbb{R}^{n\times n}$ is a nonsingular matrix, then it implies that $\boldsymbol{\Omega}_{(0)} = \boldsymbol{L}\boldsymbol{\Phi}_{(0)}$ for some matrix $\boldsymbol{L}$. Then it follows from (19) that we can write

$$\boldsymbol{\Omega}_{(k)} = \boldsymbol{\Lambda}_2^t \boldsymbol{L} \boldsymbol{\Lambda}_1^{-t} \boldsymbol{\Phi}_{(k)}. \tag{21}$$

The assumption that first n eigenvalues of $\boldsymbol{C}$ are strictly larger than the others, together with (21), implies that $\boldsymbol{\Omega}_{(k)}$ converges to zero and is asymptotically in the order of $(\lambda_{n+1}/\lambda_n)^t$ where λ_n and λ_{n+1} $(< \lambda_n)$ are nth and $(n+1)$th largest eigenvalues of $\boldsymbol{C}$.

Taking into account that $\boldsymbol{\Omega}_{(k)}$ goes to zero, the fixed point $\boldsymbol{\Phi}$ of (19) satisfies

$$\boldsymbol{\Phi} \mathsf{U}_T \left[\boldsymbol{\Phi}^\top \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}\right] = \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}. \tag{22}$$

Note that $\boldsymbol{\Lambda}_1$ is a diagonal matrix with diagonal entries λ_i for $i = 1, \ldots, n$. Thus, one can easily see that $\boldsymbol{\Phi}$ is the eigenvector matrix of $\mathsf{U}_T \left[\boldsymbol{\Phi}^\top \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}\right]$ with associated eigenvalues in $\boldsymbol{\Lambda}_1$. Note that the eigenvalues of an upper-triangular matrix are the diagonal elements. Then it follows from (22) that we have a set of equations

$$\left(\boldsymbol{\varphi}_i^\top \boldsymbol{\Lambda}_1 \boldsymbol{\varphi}_i\right) = \lambda_i, \quad i = 1, \ldots, n. \tag{23}$$

where $\boldsymbol{\varphi}_i$ is the ith column vector of $\boldsymbol{\Phi}$, i.e., $\boldsymbol{\Phi} = [\boldsymbol{\varphi}_1\, \boldsymbol{\varphi}_2 \cdots \boldsymbol{\varphi}_n]$. We can re-write (23) as

$$\sum_{i=1}^{n} \lambda_i \varphi_{ij}^2 = \lambda_j, \quad j = 1, \ldots, n, \tag{24}$$

where φ_{ij} is the (i,j)-element of $\boldsymbol{\Phi}$. Assume $n \le \mathrm{rank}(\boldsymbol{C})$, then $\lambda_i \neq 0$, $i = 1, \ldots, n$. For positive values λ_i, the only $\boldsymbol{\Phi}$ satisfying (24) is $\boldsymbol{\Phi} = \boldsymbol{\Upsilon}$. Therefore, $\boldsymbol{W} = \boldsymbol{U}_1 \boldsymbol{\Upsilon}$, implying that the fixed point of (9) is the true eigenvector matrix $\boldsymbol{U}_1$ up to a sign ambiguity. ■

References

Abed-Meraim, K., A. Chkeif, and Y. Hua: 2000, 'Fast Orthonormal PAST Algorithm'. *IEEE Signal Processing Letters* **7**(3), 60–62.

Ahn, J. H., S. Choi, and J. H. Oh: 2004, 'A New Way of PCA: Integrated-Squared-Error and EM Algorithms'. In: *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing*. Montreal, Canada.

Ahn, J. H. and J. H. Oh: 2003, 'A Constrained EM Algorithm for Principal Component Analysis'. *Neural Computation* **15**(1), 57–65.

Baldi, P. and K. Hornik: 1989, 'Neural Networks for Principal Component Analysis: Learning from Examples without Local Minima'. *Neural Networks* **2**, 53–58.

Brockett, R. W.: 1991, 'Dynamical Systems that Sort Lists, Diagonalize Matrices, and Solve Linear Programming Problems'. *Linear Algebra and Applications* **146**, 79–91.

Choi, S.: 2005, 'On Variations of Power Iteration'. In: *Proc. Int'l Conf. Artificial Neural Networks*, Vol. 2. Warsaw, Poland, pp. 145–150, Springer.

Cichocki, A. and R. Unbehauen: 1992, 'Neural Networks for Computing Eigenvalues and Eigenvectors'. *Biological Cybernetics* **68**, 155–164.

Diamantaras, K. I. and S. Y. Kung: 1996, *Principal Component Neural Networks: Theory and Applications*. John Wiely & Sons, INC.

Hua, Y., Y. Xiang, T. Chen, K. Abed-Meraim, and Y. Miao: 1999, 'A New Look at the Power Method for Fast Subspace Tracking'. *Digital Signal Processing* **9**, 297–314.

Jolliffe, I. T.: 2002, *Principal Component Analysis*. Spinger-Verlag, 2 edition.

Oja, E.: 1989, 'Neural Networks, Principal Component Analysis, and Subspaces'. *International Journal of Neural Systems* **1**, 61–68.

Roweis, S. T.: 1998, 'EM Algorithms for PCA and SPCA'. In: *Advances in Neural Information Processing Systems*, Vol. 10. pp. 626–632, MIT press.

Sanger, T. D.: 1989, 'Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network'. *Neural Networks* **2**(6), 459–473.

Tipping, M. E. and C. M. Bishop: 1999, 'Probabilistic Principal Component Analysis'. *Journal of the Royal Statistical Society B* **61**(3), 611–622.

Xu, L.: 1993, 'Least MSE Reconstruction: A Principle for Self-Organizing Nets'. *Neural Networks* **6**, 627–648.

Yang, B.: 1995, 'Projection Approximation Subsapce Tracking'. *IEEE Trans. Signal Processing* **43**(1), 95–107.