

Convergence Analysis of Complementary Candid Incremental Principal Component Analysis *

Yilu Zhang and Juyang Weng

Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{weng, zhangyil, hwangwey}@cse.msu.edu

Abstract

In this report, we analyze a proposed incremental principal component analysis algorithm, complementary candid incremental PCA algorithm, and prove that, following this algorithm, the estimated vectors $v_i(n)$ converge to $\lambda_i e_i$ when $n \rightarrow \infty$, with probability 1.

1 Introduction

Principal component analysis (PCA) is a well-known technique in data compression and feature extraction. It gives a linear transform that is used to convert a set of d -dimension data into a lower-dimensional space by minimizing the error in the least mean square (LMS) sense.

A simple approach to PCA is to solve an eigensystem problem. Given A being the sample covariance matrix of the data set, one can find its eigenvectors and eigenvalues, sort the eigenvalues in descending order, and construct a $k \times d$ matrix T with the rows being the eigenvectors corresponding to the largest k eigenvalues. T is the transform we are seeking [1]. This is the basic idea of most techniques to conduct PCA, such as QR method [2]. However, as it involves estimating the covariance matrix, the data set is usually required to be completely available at computation time. When the dimensionality of the data is high, both the computation and storage complexity grow dramatically. For example, in the eigenface method, one of the promising face recognition methods that involves PCA, a moderate grey image is of 64 rows and 88 columns, which results in a 5632-dimensional vector. Since the sample covariance matrix of a data set of d -dimensional random vectors contains $d(d+1)/2$ numbers, this amounts to 15,862,528 numbers!

Incremental principal component analysis (IPCA) techniques have been investigated and developed to compute principal component without the covariance matrix [3][4][5]. But they ran into convergence problems when facing high dimensional image vectors. We proposed a new incremental principal component analysis algorithm, complementary candid incremental PCA (CCIPCA) algorithm, that outperforms the existing methods [6], whose convergence is proved mathematically in this paper. The proof uses the concepts and the tools of stochastic approximation discussed in [7]. In the following sections, we first present the algorithm and then prove a theorem that links the algorithm with the stable solution of a differential equation. In section 4 and 5, we prove the convergence of the algorithm by satisfying the conditions of the theory.

*A technical report of computer science department at Michigan State University, MSU-CSE-01-23, August 2001.

2 The Algorithm

Without losing the generality, throughout this report, we assume that $\{u(n)\}$, a d -dimensional random vector sequence, has zero-mean and identical independent distribution. The algorithm of CCIPCA is as follows,

$$v_i(n) = \frac{n-1}{n}v_i(n-1) + \frac{1}{n}u_i(n)u_i^T(n) \frac{v_i(n-1)}{\|v_i(n-1)\|} \quad (1)$$

$$u_i(n) = u_{i-1}(n) - u_{i-1}^T(n) \frac{v_{i-1}(n)}{\|v_{i-1}(n)\|} \frac{v_{i-1}(n)}{\|v_{i-1}(n)\|} \quad (2)$$

where $i = 1, 2, \dots, k$ is the eigenvector number, n is the sample sequential number, $u_1(n) = u(n)$, and $v_i(n)$ is the estimate of the i -th dominant eigenvectors of the covariance matrix.

We going to prove that, with the algorithm given by Eq. (1) and Eq. (2), $v_i(n) \rightarrow \pm\lambda_i e_i$ when $n \rightarrow \infty$, where λ_i is the i -th largest eigenvalue of the covariance matrix of $\{u(n)\}$, and e_i is the corresponding eigenvector, under following assumptions,

1. $E\{A(n)\} = A$ for all n .
2. A is of full rank, which means λ_d , the smallest eigenvalue of A , is not zero.

3 Relation to a Differential Equation

When $i = 1$, Eq. (1) can be written as,

$$v_1(n) = v_1(n-1) + \frac{1}{n} \left(\frac{A(n)}{\|v_1(n-1)\|} - I \right) v_1(n-1) \quad (3)$$

where $A(n) = u_1(n)u_1^T(n) = u(n)u^T(n)$, or,

$$v_1(n) = v_1(n-1) + \frac{1}{n} \left(\frac{A}{\|v_1(n-1)\|} - I \right) v_1(n-1) + \frac{1}{n} \frac{A(n) - A}{\|v_1(n-1)\|} v_1(n-1) \quad (4)$$

where $A = E\{A(n)\}$ for all n .

Lemma 3.1 $\lim_{n \rightarrow \infty} P\{\sup \|\frac{A(n)-A}{\|v_1(n-1)\|} v_1(n-1)\| \geq \varepsilon\} = 0$.

Proof. Since $\|\frac{A(n)-A}{\|v_1(n-1)\|} v_1(n-1)\| = \|A(n) - A\|$ and $A = E\{A(n)\}$ for all n , it is simple to conclude

$$\lim_{n \rightarrow \infty} P\{\sup \|\frac{A(n) - A}{\|v_1(n-1)\|} v_1(n-1)\| \geq \varepsilon\} = 0$$

◁

Lemma 3.2 $v_1(n)$ is bounded with probability 1.

Proof. Let $\lambda_{max}(A(n))$ be the largest eigenvalue of $A(n)$, which is bounded w.p.1 because $A = E\{A(n)\}$ for all n and A has bounded eigenvalues.

We will consider two cases in the proof.

Case 1: If $\|v_1(n-1)\|$ is always smaller than $2\lambda_{max}(A(n))$, we know $v_1(n)$ is bounded w.p.1.

Case 2: Suppose $\|v_1(n-1)\| \geq 2\lambda_{max}(A(n))$ for certain n . According to Eq. (3), we have,

$$\begin{aligned} \|v_1(n)\|^2 &= \|v_1(n-1)\|^2 + \frac{2}{n} \frac{v_1^T(n-1)A(n)v_1(n-1)}{\|v_1(n-1)\|} - \frac{2}{n} v_1^T(n-1)v_1(n-1) \\ &\quad + \frac{1}{n^2} v_1^T(n-1)A^2(n)v_1(n-1) + \frac{1}{n^2} v_1^T(n-1)v_1(n-1) - \frac{2}{n^2} \frac{v_1^T(n-1)A(n)v_1(n-1)}{\|v_1(n-1)\|} \end{aligned} \quad (5)$$

Considering $v_1^T(n-1)A(n)v_1(n-1) \leq \lambda_{max}(A(n))v_1^T(n-1)v_1(n-1)$ ¹ and $\|v_1(n-1)\| \geq 2\lambda_{max}(A(n))$ for certain n , we have,

$$\frac{2}{n} \frac{v_1^T(n-1)A(n)v_1(n-1)}{\|v_1(n-1)\|} < \frac{1}{n} v_1^T(n-1)v_1(n-1). \quad (6)$$

Since $v_1^T(n-1)A^2(n)v_1(n-1) \leq \lambda_{max}^2(A(n))v_1^T(n-1)v_1(n-1)$, when n is large enough ($n > 2\lambda_{max}^2$), we have,

$$\frac{1}{n^2} v_1^T(n-1)A^2(n)v_1(n-1) < \frac{1}{2n} v_1^T(n-1)v_1(n-1). \quad (7)$$

Moreover, when n is large enough ($n > 2$), we have,

$$\frac{1}{n^2} v_1^T(n-1)v_1(n-1) < \frac{1}{2n} v_1^T(n-1)v_1(n-1). \quad (8)$$

Thus, from Eq. (6)-(8) and Eq. (5), we know that, for large enough n , $\|v_1(n)\| < \|v_1(n-1)\|$ and $v_1(n)$ is bounded.

We can conclude from case 1 and case 2 that, either $\|v_1(n)\| < 2\lambda_{max}(A(n))$ or $\|v_1(n)\| < \|v_1(n-1)\|$. Both cases imply that $\|v_1(n)\|$ is bounded w.p.1. \triangleleft

Theorem 3.1 *Let v_{10} be a locally asymptotically stable (in the sense of Liapunov) solution to*

$$\dot{v}_1 = \left(\frac{A}{\|v_1\|} - I \right) v_1 \quad (9)$$

with domain of attraction $\mathcal{D}(v_{10})$. If there is a compact set $\mathcal{A} \subset \mathcal{D}(v_{10})$ such that the solution of Eq. (4) satisfies $P\{v_1(n) \in \mathcal{A}\} = 1$, then $v_1(n)$ tends to v_{10} almost surely.

Proof. We prove it using Theorem 2.3.1 in Kushner and Clark [7]. Assumptions A.2.2.1, A.2.2.2, and A.2.2.3 in [7] are trivial. Lemma 3.1 fulfills the assumption of A.2.2.4. Together with Lemma 3.2, all the assumptions are satisfied and, thus, Theorem 2.3.1 in [7] implies Theorem 3.1 here. \triangleleft

Note: when using this theorem, we need to make sure ‘‘Eq. (4) satisfies $P\{v_1(n) \in \mathcal{A}\} = 1$ ’’.

¹Consider $x^T A x = (E_A x)^T \Lambda_A (E_A x)$, where the rows of E_A are eigenvectors of A , Λ_A is a diagonal matrix with the eigenvalues of A as its elements. Let $x_E = E_A x$. $x^T A x = x_E^T \Lambda_A x_E = y^T y$, where $y = x_E \Lambda_{A\sqrt{\cdot}}$ and $\Lambda_{A\sqrt{\cdot}}$ is a diagonal matrix with the square root of the eigenvalues of A as its elements. It is not difficult to see that $\|y\| \leq \lambda_{max}(A) \|x_E\|$. Since $\|E_A x\| = \|x\|$, we know $x^T A x \leq \lambda_{max}(A) \|x\|^2$. It may be proved that $x^T A^n x \leq \lambda_{max}^n(A) \|x\|^2$.

4 Prove $v_1(n) \rightarrow \pm \lambda_1 e_1$

In this section, we will first prove that the asymptotically stable solution of Eq. (9) is $\lambda_1 e_1$, and then make sure “Eq. (4) satisfies $P\{v_1(n) \in \mathcal{A}\} = 1$ ”.

Expending v_i in terms of the eigenvectors, which is a base of d -dimensional space, we have, $v_1 = \sum_{j=1}^d \alpha_j e_j$, where $\alpha_j = v_1^T e_j$. Considering $Ae_1 = \lambda_1 e_1$, we can write Eq. (9) as,

$$\dot{\alpha} = \left(\frac{A_\lambda}{\sqrt{\sum_{k=1}^d \alpha_k^2}} - I \right) \alpha$$

where, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)^T$, $\|v_1\| = \sqrt{\sum_{k=1}^d \alpha_k^2}$, and $A_\lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$.

4.1 $\|v_1\|$ is bounded

Let $r = \|v_1\| = (v_1^T v_1)^{1/2}$, we have,

$$\begin{aligned} \dot{r} &= \frac{1}{r} v_1^T \dot{v}_1 \\ &= \frac{1}{r} v_1^T \left(\frac{A}{r} - I \right) v_1 \\ &= \frac{1}{r} \left(\frac{\sum_{k=1}^d \lambda_k r^2 \cos^2 \theta_k}{r} - r^2 \right) \\ &= \sum_{k=1}^d \lambda_k \cos^2 \theta_k - r \end{aligned}$$

where θ_k is the angle between v_1 and e_k . Because $\sum_{k=1}^d \cos^2 \theta_k = 1$, we have,

$$\dot{r} = \lambda_d + \sum_{k=1}^{d-1} \cos^2 \theta_k (\lambda_k - \lambda_d) - r \quad (11)$$

Since $\lambda_1 > \lambda_2 > \dots > \lambda_d > 0$, we know $0 \leq \sum_{k=1}^{d-1} \cos^2 \theta_k (\lambda_k - \lambda_d) \leq \sum_{k=1}^{d-1} (\lambda_k - \lambda_d) < \sum_{k=1}^{d-1} \lambda_k$. The asymptotic solution of ODE $\dot{r} = c(t) - r$, where $\lambda_d \leq c(t) \leq \sum_{k=1}^d \lambda_k$, is bounded in the range of $c(t)$. Hence, $\lambda_d \leq \|v_1\| \leq \sum_{k=1}^d \lambda_k$ when $t \rightarrow \infty$.

4.2 $\alpha_j \rightarrow 0$ for $j > 1$

Let $\theta_j = \alpha_j / \alpha_1$ ($\alpha_1 \neq 0$ with probability 1 as we will see in section 4.4) for $j > 1$, and we have,

$$\begin{aligned} \dot{\theta}_j &= \frac{1}{\alpha_1^2} (\alpha_1 \dot{\alpha}_j - \alpha_j \dot{\alpha}_1) \\ &= \frac{1}{\alpha_1^2 \sum_{k=1}^d \alpha_k^2} (\lambda_j \alpha_1 \alpha_j - \lambda_1 \alpha_1 \alpha_j) \\ &= \frac{\theta_j}{\sqrt{\sum_{k=1}^d \alpha_k^2}} (\lambda_j - \lambda_1) \end{aligned} \quad (12)$$

Since $\lambda_d \leq \|v_1\| = \sqrt{\sum_{k=1}^d \alpha_k^2} < \sum_{k=1}^d \lambda_k$ and $\lambda_j < \lambda_1$, we have $\theta_j \rightarrow 0$ as $t \rightarrow \infty$. Again because of the upper bound of $\|v_1\|$, α_1 is bounded and, consequently, $\alpha_j \rightarrow 0$ for $j > 1$ when $t \rightarrow \infty$.

4.3 $\alpha_1 \rightarrow \pm\lambda_1$

From Eq. (10), we have,

$$\dot{\alpha}_1 = \left(\frac{\lambda_1}{\sqrt{\sum_{k=1}^d \alpha_k^2}} - 1 \right) \alpha_1.$$

We may drop α_j for $j > 1$ when $t \rightarrow \infty$ and get,

$$\dot{\alpha}_1 = \pm\lambda_1 - \alpha_1,$$

which means $\alpha_1 \rightarrow \pm\lambda_1$ when $t \rightarrow \infty$.

4.4 Summary

The only case in which $v_1(t)$ does not converge to $\pm\lambda_1 e_1$ is $\alpha_1(t) = 0$. When $\alpha_1(t) = 0$, $v_1(t)$ is in a $(d-1)$ -dimensional space, $R_1^{d-1} = \text{span}\{e_i, i = 2, \dots, d\}$. In other words, the domain of attraction of $\{\pm\lambda_1 e_1\}$ ($DA(\{\pm\lambda_1 e_1\})$) is $R^d - R_1^{d-1}$. Since it is very unlikely that we choose $v_1(0)$ in the subspace R_1^{d-1} , $v_1(n)$ enters $DA(\{\pm\lambda_1 e_1\})$ with probability one. Applying Theorem 3.1, we have $v_1(n) \rightarrow \pm\lambda_1 e_1$ with probability 1 when $n \rightarrow \infty$.

5 Prove $v_i(n) \rightarrow \pm\lambda_i e_i$ with Induction

We want to prove that $v_i(n) \rightarrow \pm\lambda_i e_i$ under induction assumption that $v_j(n) \rightarrow \pm\lambda_j e_j$ ($j < i$).

From Eq. (2), we have,

$$\begin{aligned} u_i(n) &= u_{i-1}(n) - \frac{v_{i-1}(n)}{\|v_{i-1}(n)\|} \left[\frac{v_{i-1}^T(n)}{\|v_{i-1}(n)\|} u_{i-1}(n) \right] \\ &= \prod_{j=1}^{i-1} \left[I - \frac{v_j(n)v_j^T(n)}{\|v_j(n)\|^2} \right] u_1(n) \end{aligned} \quad (15)$$

To simply notation, we define,

$$\Pi_i(n) = \prod_{j=1}^{i-1} \left[I - \frac{v_j(n)v_j^T(n)}{\|v_j(n)\|^2} \right],$$

where $\Pi_i(n) = I$ if the superscript and the subscript cross over. So, Eq. (1) can be written as,

$$v_i(n) = v_i(n-1) + \frac{1}{n} \left(\frac{\Pi_i(n)A(n)\Pi_i(n)}{\|v_i(n-1)\|} - I \right) v_i(n-1) \quad (17)$$

where $A(n) = u_1(n)u_1(n)$.

Similar to the case of $v_1(n)$, we may find the related differential equation of Eq. (17) as,

$$\dot{v}_i = \left(\frac{\Pi_i A \Pi_i}{\|v_i\|} - I \right) v_i \quad (18)$$

where $A = E\{A(n)\}$ and $\Pi_i = \Pi_{j=1}^{i-1} \left[I - \frac{v_j v_j^T}{\|v_j\|^2} \right]$.

With the induction assumption, we write, for $j < i$,

$$\frac{v_j}{\|v_j\|} = e_j + \varepsilon_j w_j \quad (19)$$

where w_j is a time-variable unit-length vector, and for $j < i$, $\varepsilon_j(t) \rightarrow 0$ as $t \rightarrow \infty$. (Following analysis will be similar if we write, $\frac{v_j}{\|v_j\|} = -e_j + \varepsilon_j w_j$.)

From the definition of Π_i we have,

$$\begin{aligned} \Pi_i(t) &= \Pi_{j=1}^{i-1} [I - (e_j + \varepsilon_j(t)w_j(t))(e_j + \varepsilon_j(t)w_j(t))^T] \\ &= \Pi_{j=1}^{i-1} [I - e_j e_j^T - \varepsilon_j(t)(w_j(t)e_j^T + e_j w_j(t)^T + w_j(t)w_j(t)^T)] \\ &= I - \sum_{j=1}^{i-1} e_j e_j^T - O(\varepsilon(t)) \end{aligned} \quad (20)$$

where $\varepsilon(t) = \arg \max_{j < i} \varepsilon_j(t)$. Further, we have,

$$\begin{aligned} \Pi_i(t) A \Pi_i^T(t) &= \Pi_i(t) [A - \sum_{j=1}^{i-1} \lambda_j e_j e_j^T - O(\varepsilon(t))] \\ &= A - \sum_{j=1}^{i-1} \lambda_j e_j e_j^T - O(\varepsilon(t)) \end{aligned} \quad (21)$$

Expanding v_i in terms of the eigenvectors, we have,

$$v_i(t) = \sum_{k=1}^d \alpha_k(t) e_k, \quad (22)$$

where $\alpha_k(t) = v_i^T(t) e_k$. Substituting Eq. (21) and Eq. (22) into Eq. (18), and ignoring $O(\varepsilon(t))$ when t is large, we have,

$$\dot{\alpha} = \left(\frac{A_{\lambda_i}}{\sqrt{\sum_{k=1}^d \alpha_k^2}} - I \right) \alpha$$

where, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)^T$, $\|v_i\| = \sqrt{\sum_{k=1}^d \alpha_k^2}$, and $A_{\lambda_i} = \text{diag}(0, 0, \dots, 0, \lambda_i, \lambda_{i+1}, \dots, \lambda_d)$. Since $\dot{\alpha}_j = -\alpha_j$ for $j < i$, we have $\alpha_j \rightarrow 0$ when $t \rightarrow \infty$. Dropping α_j ($j < i$), we have very similar differential equations as in Eq. (10). Following the proof in section 4.1–4.3, we can conclude with $\alpha_i \rightarrow \pm \lambda_i$ and $\alpha_j \rightarrow 0$ ($j \neq i$) when $t \rightarrow \infty$.

Similar to the analysis in section 4.4, we have $v_i(n) \rightarrow \pm \lambda_i e_i$ with probability one when $n \rightarrow \infty$.

6 Conclusions

As a conclusion of above analysis, with the algorithm given by Eq. (1) and Eq. (2), $v_i(n) \rightarrow \pm \lambda_i e_i$ with probability one when $n \rightarrow \infty$, where λ_i is the i -th largest eigenvalue of the covariance matrix of $\{u(n)\}$, and e_i is the corresponding eigenvector.

References

- [1] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, NY, second edition, 1990.
- [2] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
- [3] E. Oja, *Subspace Methods of Pattern Recognition*, Research Studies Press, Letchworth, UK, 1983.
- [4] E. Oja and J. Karhunen, “On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix,” *Journal of Mathematical Analysis and Application*, vol. 106, pp. 69–84, 1985.
- [5] T.D. Sanger, “Optimal unsupervised learning in a single-layer linear feedforward neural network,” *IEEE Trans. Neural Networks*, vol. 2, pp. 459–473, 1989.
- [6] Y. Zhang and J. Weng, “Complementary candid incremental principal component analysis,” Tech. Rep. MSU-CSE-01-24, Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, August 2001.
- [7] H.J. Kushner and D.S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*, Springer-Verlag, New York, 1978.