CONVERGENCE STUDY OF PRINCIPAL COMPONENT ANALYSIS ALGORITHMS

CHANCHAL CHATTERJEE
NEWPORT CORPORATION, 1791 DEERE AVENUE, IRVINE, CA 92606
VWANI P. ROYCHOWDHURY
ELECTRICAL ENGINEERING DEPARTMENT, UCLA, LOS ANGELES, CA 90095
AND
EDWIN K.P. CHONG
SCHOOL OF ELECT. AND COMPUTER ENGG., PURDUE UNIVERSITY, W. LAFAYETTE, IN 47907

Abstract - We investigate the convergence properties of two different principal component analysis algorithms, and analytically explain some commonly observed experimental results. We use two different methodologies to analyze the two algorithms. The first methodology uses the fact that both algorithms are stochastic approximation procedures. We use the theory of stochastic approximation, in particular the results of Fabian, to analyze the *asymptotic mean square errors* (AMSEs) of the algorithms. This analysis reveals the conditions under which the algorithms produce smaller AMSEs, and also the conditions under which one algorithm has a smaller AMSE than the other. We next analyze the *asymptotic mean errors* (AMEs) of the two algorithms in the neighborhood of the solution. This analysis establishes the conditions under which the AMEs of the minor eigenvectors go to zero faster. Furthermore, the analysis makes explicit that increasing the gain parameter up to an upper bound improves the convergence of all eigenvectors. We also show that the AME of one algorithm goes to zero faster than the other. Experiments with multi-dimensional Gaussian data corroborate the analytical findings presented here.

1. Introduction

We investigate the convergence properties of different algorithms for principal component analysis (PCA). Existing literature on PCA (see [1]) reveal a number of algorithms that are derived from: (i) anti-Hebbian learning [1,5], (ii) Hebbian learning [1,5,8,9], (iii) lateral interaction algorithms [1,5], and (iv) gradient-based learning [1,5,9]. Since there are multiple algorithms for PCA, and each one requires a different amount of computation at each update, yet leads to the same set of final results, we are led to ask the question: "which algorithm converges faster than others ?".

Among the many algorithms for PCA, we focus on two algorithms that are commonly used in most applications. It can be shown [1,5] that several other algorithms for PCA are related to these basic procedures. In both of these algorithms, we are given a sequence of random matrices $\{A_k \in \Re^{dXd}\}$, with $\lim_{k\to\infty} E[A_k]=A$, where A_k represents an online observation of the application. Note that one common realization of A_k is from a sequence of random vectors $\{\mathbf{x}_k\}$ as $A_k=\mathbf{x}_k\mathbf{x}_k^T$. For each sample A_k, we compute a matrix $W_k \in \Re^{dXp}$ ($p \le d$) by these algorithms, such that W_k tends, with probability one (w.p.1), to a matrix Φ_p whose columns are the p ($\le d$) eigenvectors of A ordered by decreasing eigenvalue. Conforming to existing literature, we refer to Φ_p as the *principal eigenvector matrix* of A. The first column of Φ_p is the first principal eigenvector of A, and so on.

The first algorithm that we consider is due to Oja and Karhunen [6], and Sanger [8], and can be derived from Hebbian learning with a multi-unit network [1]. This algorithm is

$$W_{k+1} = W_k + \eta_k\left(A_k W_k - W_k \mathrm{UT}_\gamma\left[W_k^T A_k W_k\right]\right), \quad (1)$$

where $\{\eta_k\}$ is a sequence of scalar gains. Here $\mathrm{UT}_\gamma[\cdot]$ sets all elements below the diagonal of its matrix argument to zero, and multiplies all elements above the diagonal with γ (≥ 1).

The second algorithm is due to Xu [9], and can be derived from a least mean square error criterion. The algorithm is

$$W_{k+1} = W_k + \eta_k\Big(2A_k W_k - W_k \mathrm{UT}_\gamma\left[W_k^T A_k W_k\right] - A_k W_k \mathrm{UT}_\gamma\left[W_k^T W_k\right]\Big). \quad (2)$$

Note that algorithm (2) uses nearly twice as much computation than algorithm (1) for each update of W_k. Hence, it is natural to ask the question: "what benefits can

0-7803-4122-8/97 $10.00©1997 IEEE

we derive from algorithm (2) at the cost of higher computation ?".

In our analyses, we use two different methodologies to answer the above-mentioned question. The first methodology uses the fact that both algorithms (1) and (2) are stochastic approximation procedures [3,5]. We use the theory of stochastic approximation, in particular the results of Fabian [3], to analyze the *asymptotic mean square errors* (AMSEs) for the algorithms. This analysis reveals the following facts: (i) by using the parameter $\gamma>1$ both algorithms lead to a larger AMSE than $\gamma=1$ for the minor eigenvectors; and (ii) by using the parameter $\gamma>1$, algorithm (2) leads to a smaller AMSE than algorithm (1) for the minor eigenvectors.

We next perform a convergence analysis of the *asymptotic mean errors* (AMEs) of the two algorithms, with $\gamma=1$ and $\gamma>1$, in the neighborhood of the solution. This analysis makes explicit the following facts for both algorithms: (i) with $\gamma>1$ the AMEs of the minor eigenvectors go to zero faster than with $\gamma=1$, (ii) with increasing $\{\eta_k\}$, up to an upper bound, we improve the convergence of all eigenvectors, and (iii) the AME of algorithm (2) goes to zero faster than algorithm (1).

In summary, the above results reveal the following facts: (i) although both algorithms converge faster to the solution for $\gamma>1$, they produce a larger AMSE, and (ii) by choosing $\gamma>1$, algorithm (2) converges faster than algorithm (1), and leads to a smaller AMSE of the estimates. Thus, a clear trade-off between computation and convergence is made explicit. By using algorithm (2) with $\gamma>1$, we perform more computations, but converge faster than algorithm (1), and also produce a smaller AMSE of the estimates when compared to algorithm (1).

In section 2 we provide a mathematical background of the PCA algorithms. In Section 3, we compare the asymptotic mean square errors of the estimated eigenvectors. In Section 4, we compare the asymptotic mean errors of the estimated eigenvectors. Section 5 has the experimental results.

2. Background and Definitions

In this section, we provide a mathematical background of the PCA algorithms, and give the definitions that are used here. We are given a sequence $\{A_k \in \Re^{dXd}\}$, with $\lim_{k\to\infty} E[A_k|A_0,...,A_{k-1}]=A$. The *p principal eigenvectors* of A are the p eigenvectors of A corresponding to the p largest eigenvalues. In the following discussion, we denote $\lambda_1>\lambda_2>...>\lambda_p\geq...\geq\lambda_d>0$ as the eigenvalues of A, and ϕ_i as the eigenvector corresponding to λ_i such that ϕ_1, ..., ϕ_d are orthonormal. Let $\Phi=[\phi_1...\phi_d]$ and $\Lambda=\mathrm{diag}(\lambda_1,...,\lambda_d)$ denote the matrix of eigenvectors and eigenvalues of A. Note that if ϕ_i is an eigenvector, then $a_i\phi_i$ for $a_i=\pm1$ is also an eigenvector.

Let the i^{th} column of W_k be $\mathbf{w}_k^i$. Then, algorithm (1) can be written as

$$\mathbf{w}_{k+1}^i = \mathbf{w}_k^i + \eta_k\Big(A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{iT}A_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{jT}A_k\mathbf{w}_k^i\Big) \text{ for } i=1,...,p. \quad (3)$$

Here parameter γ is greater than or equal to one. Algorithm (2) can be similarly written as

$$\mathbf{w}_{k+1}^i = \mathbf{w}_k^i + \eta_k\Big(2A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{iT}A_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{jT}A_k\mathbf{w}_k^i - A_k\mathbf{w}_k^i\mathbf{w}_k^{iT}\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}A_k\mathbf{w}_k^j\mathbf{w}_k^{jT}\mathbf{w}_k^i\Big) \text{ for } i=1,...,p. \quad (4)$$

Since it is clear that we require more computation for (4) than (3) at each update, it is necessary to determine the relative convergence rates to compare the benefits at the cost of computation. For both algorithms, we have the following assumptions and propositions.

Assumption (A1). Each A_k is bounded with probability one, symmetric, real, nonnegative definite, and $\lim_{k\to\infty} E[A_k|A_0,...,A_{k-1}]=A$, where A is positive definite.

Assumption (A2). $\{\eta_k\in\Re^+\}$ is a decreasing sequence, with $\sum_{k=0}^{\infty}\eta_k=\infty$, $\sum_{k=0}^{\infty}\eta_k^r<\infty$ for some $r>1$, and $\lim_{k\to\infty}\sup(\eta_k^{-1}-\eta_{k-1}^{-1})<\infty$.

Assumption (A3). The p largest eigenvalues of A are positive and each of unit multiplicity.

Proposition (P1). For algorithms (3) and (4) let assumptions A1-A3 hold. Then $\mathbf{w}_k^i$ tends either to ϕ_i or $-\phi_i$ with probability one as $k\to\infty$. ■

Proposition (P2). For algorithms (3) and (4) let assumptions A1-A3 hold. Assume that $\mathbf{w}_0^i$ is bounded. Then, there exists a uniform upper bound of η_k such that $\mathbf{w}_k^i$ is uniformly bounded w.p.1. ■

3. Comparison of the Asymptotic Mean Squared Errors

In order to compare the asymptotic mean squared errors of the two algorithms, we use the stochastic approximation

results of Fabian [3]. Fabian established rates of convergence with $\eta_k=\delta k^{-\alpha}$ for $\frac{1}{2}<\alpha\le 1$ and $\delta>0$. In the following Theorem, we specialize Fabian's results to suit the present analysis.

Theorem 1 (Fabian[3]). *Let $\Im_k$ be a non-decreasing sequence of σ-fields. Suppose $\mathbf{e}_k$, $\mathbf{v}_k \in \Re^d$, $\Gamma_k \in \Re^{dXd}$, Σ, Γ, $P \in \Re^{dXd}$, Γ is positive definite, P is orthonormal and $P^T\Gamma P = \Theta$ is diagonal, where $\Theta = \mathrm{diag}(\theta_1,...,\theta_d)$. Suppose Γ_k and V_{k-1} are $\Im_k$-measurable, $\alpha, \beta \in \Re$ and*

$$\Gamma_k \to \Gamma, \quad E_{\Im_k}[\mathbf{v}_k]=\mathbf{0}, \text{ and } E_{\Im_k}[\mathbf{v}_k\mathbf{v}_k^T]\to\Sigma \quad \text{as } k\to\infty. \tag{5}$$

Suppose that $\theta=\min\{\theta_1,...,\theta_d\}$, $\beta_+=\beta$ if $\alpha=1$, $\beta_+=0$ if $\alpha\neq 1$,

$$0<\alpha\le 1,\ \beta\ge 0,\ \beta_+<2\theta \tag{6}$$

and

$$\mathbf{e}_{k+1}=\mathbf{e}_k-k^{-\alpha}\Gamma_k\mathbf{e}_k+k^{-(\alpha+\beta)/2}\mathbf{v}_k. \tag{7}$$

Then the asymptotic distribution of $k^{\beta/2}\mathbf{e}_k$ is Gaussian with mean $\mathbf{0}$ and covariance PMP^T where

$$M_{ij}=\left(P^T\Sigma P\right)_{ij}\left(\theta_i+\theta_j-\beta_+\right)^{-1}. \tag{8}$$

■

Outline of Approach and Summary of Results. In the following two sections, we first present each algorithm (i.e., algorithms (1) and (2)) in the format of eqn. (7). We next evaluate the parameters given in Theorem 1, and finally, compute the asymptotic mean squared errors for each algorithm. Based on the analyses, our primary results can be summarized in the theorem below.

Theorem 2. *For both algorithms (1) and (2), the AMSEs for the minor eigenvectors increase for larger values of γ, with the smallest AMSE for γ=1. Moreover, both algorithms (1) and (2) lead to the same AMSE for γ=1 for all eigenvectors. However, if $\gamma>1$, then algorithm (2) leads to a smaller AMSE than algorithm (1) for the minor eigenvectors.*

Proof. Proof of theorem follows directly from the analyses in the next section (see (18) and (20)). ■

3.1 Analysis for Algorithm (1)

Step 1: Formulation of Algorithm (1) as Eqn. (7). We now analyze algorithm (1) in the framework of eqn. (7) in Theorem 1. In the representation of algorithm (1) as given in (3), we have $\eta_k = k^{-\alpha}$ for $\frac{1}{2} < \alpha \le 1$. Notice that this choice of η_k satisfies assumption A2. We define

$$\mathbf{h}(\mathbf{w}_k^i, A_k) = A_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i\,T}A_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j\,T}A_k\mathbf{w}_k^i. \tag{9}$$

Let $\Im_k$ be a σ-algebra generated by $A_0,...,A_{k-1}$. Clearly, $\Im_1 \subset \Im_2 \subset \ldots \subset \Im_k$, and hence, $\Im_k$ is non-decreasing. Furthermore, let $E_{\Im_k}[A_k]=\bar{A}_k$ such that as $k\to\infty$, $\bar{A}_k \to A$ w.p.1. Define

$$\bar{\mathbf{h}}(\mathbf{w}_k^i) = E_{\Im_k}[\mathbf{h}(\mathbf{w}_k^i, A_k)]$$
$$= \bar{A}_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i\,T}\bar{A}_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j\,T}\bar{A}_k\mathbf{w}_k^i. \tag{10}$$

Let $\mathbf{e}_k^i = \mathbf{w}_k^i - a_i\phi_i$ (a_i=±1) be the error of $\mathbf{w}_k^i$ from its asymptotically stable point $a_i\phi_i$. Then, by using the Mean-Value theorem, and $\eta_k = k^{-\alpha}$, we obtain from (10) and (3)

$$\mathbf{e}_{k+1}^i = \mathbf{e}_k^i + k^{-\alpha}\frac{d\bar{\mathbf{h}}(\overline{\mathbf{w}}_k^i)}{d\mathbf{w}_k^i}\mathbf{e}_k^i$$
$$+k^{-(\alpha+\beta)/2}k^{-(\alpha-\beta)/2}\left(\mathbf{h}(\mathbf{w}_k^i, A_k)-\bar{\mathbf{h}}(\mathbf{w}_k^i)\right), \tag{11}$$

where $\overline{\mathbf{w}}_k^i$ is on the segment of $\mathbf{w}_k^i$ and $a_i\phi_i$ for a_i=±1. Comparing (11) with (7), we obtain

$$\Gamma_k^i = -\frac{d\bar{\mathbf{h}}(\overline{\mathbf{w}}_k^i)}{d\mathbf{w}_k^i} = \overline{\mathbf{w}}_k^{i\,T}\bar{A}_k\overline{\mathbf{w}}_k^i I + 2\overline{\mathbf{w}}_k^i\overline{\mathbf{w}}_k^{i\,T}\bar{A}_k$$
$$+\gamma\sum_{j=1}^{i-1}\overline{\mathbf{w}}_k^j\overline{\mathbf{w}}_k^{j\,T}\bar{A}_k - \bar{A}_k, \tag{12}$$

and

$$\mathbf{v}_k^i = k^{-(\alpha-\beta)/2}\left(\mathbf{h}(\mathbf{w}_k^i, A_k)-\bar{\mathbf{h}}(\mathbf{w}_k^i)\right).$$

Step 2: Optimal Choices of α, β, β_+^i, and θ^i. From (11), we see that for $E_{\Im_k}[\mathbf{v}_k^i\mathbf{v}_k^{i\,T}]$ to converge, a sufficient condition is $\alpha \ge \beta$. Furthermore, we need β to be as large as possible, in order to obtain the highest rate of convergence, since the rate of convergence is proportional to $k^{\beta/2}$ by Theorem 1. Since $\alpha \le 1$, we obtain the "best rate" of convergence for $\alpha=\beta=1$.

Since by Proposition P1, we have $\mathbf{w}_k^i \to \pm\phi_i$ w.p.1 as $k \to \infty$, we obtain from (12)

$$\Gamma_k^i \xrightarrow{k} \Gamma_i = \lambda_i I + 2\lambda_i\phi_i\phi_i^T + \gamma\sum_{j=1}^{i-1}\lambda_j\phi_j\phi_j^T - A. \tag{13}$$

We see from (13) that the eigenvectors of Γ_i are ϕ_1, ϕ_2, ..., ϕ_d, and thus, $P_i = \Phi$. The eigenvalues of Γ_i are given by

$$\Gamma_i\phi_q = \begin{cases} (\lambda_i + (\gamma-1)\lambda_q)\phi_q & \text{for } q < i \\ 2\lambda_i\phi_q & \text{for } q = i. \\ (\lambda_i - \lambda_q)\phi_q & \text{for } q > i \end{cases} \quad (14)$$

Hence, Θ_i is given by (we assume $i > 1$, and the results for $i = 1$ are similar)

$$\Theta_i = \text{diag}\big(\lambda_i + (\gamma-1)\lambda_1,...,\lambda_i + (\gamma-1)\lambda_{i-1}, 2\lambda_i, \lambda_i - \lambda_{i+1},...,\lambda_i - \lambda_d\big). \quad (15)$$

Thus,

$$\theta^i = \min\{\lambda_i + (\gamma-1)\lambda_1,...,\lambda_i + (\gamma-1)\lambda_{i-1}, 2\lambda_i, \lambda_i - \lambda_{i+1},...,\lambda_i - \lambda_d\} = \lambda_i - \lambda_{i+1}$$

and $\beta_+^i < 2\theta^i = 2(\lambda_i - \lambda_{i+1})$. This gives us two conditions for the choice of α, β and β_+^i as below

Condition 1:

If $(\lambda_i - \lambda_{i+1}) \le ½$ then choose $\alpha = \beta \ne 1$, and $\beta_+^i = 0$.

Condition 2:

If $(\lambda_i - \lambda_{i+1}) > ½$ then choose $\alpha = \beta = 1$, and $\beta_+^i = 1$.

Step 3: Computation of the Asymptotic Mean Squared Error (AMSE). By using Proposition P1 and assumption A1, we obtain the diagonal elements M_{jj}^i for j=1,...,d, of the covariance matrix M_i of the asymptotic Gaussian distribution as

$$M_{jj}^i = \begin{cases} (\Phi^T\Sigma_i\Phi)_{jj}(2(\lambda_i + (\gamma-1)\lambda_j) - \beta_+^i)^{-1} & \text{for } j < i \\ (\Phi^T\Sigma_i\Phi)_{jj}(4\lambda_i - \beta_+^i)^{-1} & \text{for } j = i \\ (\Phi^T\Sigma_i\Phi)_{jj}(2(\lambda_i - \lambda_j) - \beta_+^i)^{-1} & \text{for } j > i \end{cases} \quad (16)$$

The asymptotic mean squared error of algorithm (1) is given by tr[M_i]. In order to compute this, we define $k_j = (\Phi^T\Sigma_i\Phi)_{jj}$. If we assume that $A_k = \mathbf{x}_k\mathbf{x}_k^T$, and that $\mathbf{x}_k \in \Re^d$ is from an asymptotically Gaussian distribution with zero mean and covariance A, then we obtain

$$k_j = (\Phi^T\Sigma_i\Phi)_{jj} = \begin{cases} (1-\gamma)^2\lambda_i\lambda_j & \text{for } j < i \\ 0 & \text{for } j = i. \\ \lambda_i\lambda_j & \text{for } j > i \end{cases} \quad (17)$$

The asymptotic mean squared error (AMSE) for algorithm (1) is now given by

$$AMSE = \text{tr}[M_i] = \sum_{j=1}^{i-1} k_j\left(2\lambda_i + 2(\gamma-1)\lambda_j - \beta_+^i\right)^{-1} + \sum_{j=i+1}^{d} k_j\left(2\lambda_i - 2\lambda_j - \beta_+^i\right)^{-1} \quad \text{for } i=2,...,d, \quad (18)$$

where the k_j's are obtained from (17).

3.2 Analysis for Algorithm (2)

Step 1: Formulation of Algorithm (2) as Eqn. (7). We now repeat the above analysis for algorithm (2). From (4), we obtain

$$\Gamma_k^i \xrightarrow{k} \Gamma_i = \lambda_i I + 4\lambda_i\phi_i\phi_i^T + 2\gamma\sum_{j=1}^{i-1}\lambda_j\phi_j\phi_j^T - A. \quad (19)$$

As in algorithm (1), the eigenvectors of Γ_i are ϕ_1, ϕ_2, ..., ϕ_d, and thus, $P_i = \Phi$. Furthermore, $\theta_i = \lambda_i - \lambda_{i+1}$, and $\beta_+^i < 2\theta_i = 2(\lambda_i - \lambda_{i+1})$. The two conditions for the choice of α, β and β_+^i as given in Section 3.1 are still valid.

Step 3: Computation of the Asymptotic Mean Squared Error (AMSE). We obtain from (4)

$$AMSE = \text{tr}[M_i] = \sum_{j=1}^{i-1} k_j\left(2\lambda_i + 2(2\gamma-1)\lambda_j - \beta_+^i\right)^{-1} + \sum_{j=i+1}^{d} k_j\left(2\lambda_i - 2\lambda_j - \beta_+^i\right)^{-1} \quad \text{for } i=2,...,d, \quad (20)$$

where the k_j's are obtained from (17).

4. Comparison of the Asymptotic Mean Errors

In this section, we study the convergence of the expectations of the errors in the neighborhood of the solution. The main results are summarized in Theorem 3 in Section 4.3.

4.1 Analysis for Algorithm (1)

Similar to Section 3.1, let $\Im_k$ be a σ-algebra generated by $A_0,...,A_{k-1}$, and let $E_{\Im_k}[A_k] = \bar{A}_k$ such that as $k \to \infty$, $\bar{A}_k \to A$ w.p.1. From (3), we obtain

$$E_{\Im_k}\left[\mathbf{w}_{k+1}^i\right] = \mathbf{w}_k^i + \eta_k\left(\bar{A}_k\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i\,T}\bar{A}_k\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j\,T}\bar{A}_k\mathbf{w}_k^i\right) \quad \text{for } i = 1,...,p. \quad (21)$$

Then, for large k we have

$$E_{\Im_k}\left[\mathbf{w}_{k+1}^i\right] = \mathbf{w}_k^i + \eta_k\left(A\mathbf{w}_k^i - \mathbf{w}_k^i\mathbf{w}_k^{i\,T}A\mathbf{w}_k^i - \gamma\sum_{j=1}^{i-1}\mathbf{w}_k^j\mathbf{w}_k^{j\,T}A\mathbf{w}_k^i\right) \quad \text{for } i = 1,...,p. \quad (22)$$

Let $\mathbf{e}_k^i = \mathbf{w}_k^i - a_i\phi_i$ (a_i=±1) be the error of $\mathbf{w}_k^i$ from its asymptotically stable point $a_i\phi_i$. We have from (22)

$$E_{\Im_k}\left[\mathbf{e}_{k+1}^i\right] = H_{ii}\mathbf{e}_k^i + \sum_{j=1}^{i-1} H_{ji}\mathbf{e}_k^j - \eta_k \mathbf{g}_i(\mathbf{e}_k^1,...,\mathbf{e}_k^i) \quad \text{for } i = 1,...,p. \tag{23}$$

with

$$H_{ii} = I + \eta_k\left(A - \lambda_i I - 2\lambda_i\phi_i\phi_i^T - \gamma\sum_{j=1}^{i-1}\lambda_j\phi_j\phi_j^T\right)$$

for i=1,...,p and $H_{ji} = -\eta_k\gamma a_j a_i \lambda_i \phi_j\phi_i^T$ for j=1,...,i–1. By representing (23) in a matrix form, we obtain the following matrix equation

$$E_{\Im_k}\begin{bmatrix}\mathbf{e}_{k+1}^1\\ \mathbf{e}_{k+1}^2\\ ...\\ \mathbf{e}_{k+1}^p\end{bmatrix} = \begin{bmatrix}H_{11} & 0 & ... & 0\\ H_{21} & H_{22} & ... & 0\\ ... & ... & ... & ...\\ H_{p1} & H_{p2} & ... & H_{pp}\end{bmatrix}\begin{bmatrix}\mathbf{e}_k^1\\ \mathbf{e}_k^2\\ ...\\ \mathbf{e}_k^p\end{bmatrix} - \eta_k\begin{bmatrix}\mathbf{g}_1\\ \mathbf{g}_2\\ ...\\ \mathbf{g}_p\end{bmatrix}. \tag{24}$$

The matrix H=$[H_{ij}]\in\Re^{dp\text{X}dp}$ is of lower triangular block form, whose diagonal blocks are the matrices $H_{11},...,H_{pp}$. Define $\mathbf{e}_k^T = [\mathbf{e}_k^{1T},...,\mathbf{e}_k^{pT}] \in \Re^{dp}$ and $\mathbf{g}^T = [\mathbf{g}_1^T,...,\mathbf{g}_p^T] \in \Re^{dp}$. Taking the expectations of both sides of (24), we obtain:

$$E[\mathbf{e}_{k+1}] = HE[\mathbf{e}_k] - \eta_k E[\mathbf{g}]. \tag{25}$$

The eigenvalues of H are the eigenvalues of the diagonal blocks of $H_{11},...,H_{pp}$. Each of these matrices have the same eigenvectors $\phi_1,...,\phi_d$. The eigenvalues can be obtained from the following expressions

$$H_{ii}\phi_q = \begin{cases}\left(1-\eta_k\left(\lambda_i + (\gamma-1)\lambda_q\right)\right)\phi_q & \text{for } q < i\\ (1-2\eta_k\lambda_i)\phi_q & \text{for } q = i.\\ \left(1-\eta_k\left(\lambda_i - \lambda_q\right)\right)\phi_q & \text{for } q > i\end{cases} \tag{26}$$

By choosing η_k within an upper bound according to Proposition P2, and due to assumption A3, we observe that all eigenvalues of H_{ii} are within the range (0,1). Furthermore, the vector $\mathbf{g}$ has the properties $\mathbf{g}(\mathbf{0}) = \mathbf{0}$ and $\partial\mathbf{g}(\mathbf{0})/\partial\mathbf{e}_k = 0$. Therefore, in the neighborhood of zero, we can ignore $\mathbf{g}$. Thus, following (25), $E[\mathbf{e}_k]$ goes to zero.

4.2 Analysis for Algorithm (2)

We now repeat the above analysis for algorithm (2). From (4), we obtain an equation same as (23) with

$$H_{ii} = I + \eta_k\left(A - \lambda_i I - 4\lambda_i\phi_i\phi_i^T - 2\gamma\sum_{j=1}^{i-1}\lambda_j\phi_j\phi_j^T\right) \quad \text{for } i = 1,...,p. \tag{27}$$

The eigenvalues of H_{ii} are obtained as follows:

$$H_{ii}\phi_q = \begin{cases}\left(1-\eta_k\left(\lambda_i + (2\gamma-1)\lambda_q\right)\right)\phi_q & \text{for } q < i\\ (1-4\eta_k\lambda_i)\phi_q & \text{for } q = i.\\ \left(1-\eta_k\left(\lambda_i - \lambda_q\right)\right)\phi_q & \text{for } q > i\end{cases} \tag{28}$$

4.3 Speed of Convergence

Theorem 3. *For algorithms (1) and (2), assuming $\gamma > 1$, the following hold for the matrix H_{ii} for each i:*

(1) The eigenvectors of H_{ii} are $\phi_1,...,\phi_d$, i.e., the eigenvectors of A.

(2) The eigenvalues corresponding to $\phi_1,...,\phi_{i-1}$ decrease as γ increases, and the remaining eigenvalues are independent of γ.

(3) All eigenvalues decrease as η_k increases.

Proof. The proof is a direct result of (26) and (28). ■

Since the speed of convergence is determined by the eigenvalues of H=$[H_{ij}]$, by the above analysis and Theorem 3, we conclude the following for algorithms (1) and (2), in the neighborhood of the solution, and ignoring $\mathbf{g}$:

(1) For $i > 1$, $\|E[\mathbf{e}_k^i]\|$ goes to zero faster asymptotically as γ increases.

(2) For all i, $\|E[\mathbf{e}_k^i]\|$ goes to zero faster as η_k increases up to an upper bound determined by Proposition P2.

(3) For $i > 1$, $\|E[\mathbf{e}_k^i]\|$ goes to zero faster asymptotically for algorithm (2) than for algorithm (1).

Clearly, the mean errors for the minor components go to zero faster asymptotically as γ or η_k increases, and also for algorithm (2) than for algorithm (1). These observations are further illustrated in the experimental results.

5. Experimental Results

We generate ten-dimensional Gaussian data from the first covariance matrix in [7], with the covariance matrix multiplied by twenty (see [2]) for data in $\Re^{10}$. The ten eigenvalues of the covariance matrix A are 117.996, 55.644, 34.175, 20.589, 7.873, 5.878, 1.743, 1.423, 1.213, and 1.007.

We generated 1000 samples of zero-mean multivariate Gaussian data in $\Re^{10}$. Let the k^{th} sample be denoted by $\mathbf{x}_k$. We computed A_k as $\mathbf{x}_k\mathbf{x}_k^T$, and then obtained an estimate W_{1000} of the eigenvector matrix Φ by algorithms 1 and 2 for all 1000 samples. We refer to one application of each

algorithm for one sample A_k as *one iteration* of the algorithm. We first computed the AMSEs for the first four principal eigenvectors by collecting the results for the last 20 iterations out of a total 1000 iterations. The results of the experiment are shown in Table 1.

Table 1. Estimated AMSE for algorithms (1) and (2) for multivariate Gaussian data.

γ	ϕ_1		ϕ_2	
	Alg. (1)	Alg. (2)	Alg. (1)	Alg. (2)
1.0	0.00211	0.00201	0.00128	0.00125
1.5	0.00211	0.00201	0.00196	0.00160
2.0	0.00211	0.00201	0.00198	0.00166
2.5	0.00211	0.00201	0.00215	0.00175
3.0	0.00211	0.00201	0.00298	0.00185

Table 1 reflects the following results stated in Theorem 2. We observe that: (1) the smallest AMSE is for γ=1 for all eigenvectors; (2) the AMSEs increase as γ increases for the minor eigenvectors; and (3) algorithm 2 leads to a smaller AMSE than algorithm 1 for the minor eigenvectors for γ>1.

We next present the results for AMEs as given in Section 4.3. In order to estimate the error for each eigenvector, we compute the direction cosine given by

$$\text{Direction Cosine} = \left|\mathbf{w}_k^{i\,T}\phi_i\right| \Big/ \left\|\mathbf{w}_k^i\right\|\left\|\phi_i\right\|,$$

where $\mathbf{w}_k^i$ is the estimated i^{th} principal eigenvector at k^{th} iteration of the adaptive algorithms, and ϕ_i is the actual i^{th} principal eigenvector computed from all collected samples by a standard method [4]. Note that the maximum value of the direction cosine is one when $\mathbf{w}_k^i$ is exactly same as ϕ_i.

Figure 1 shows the iterates of the first four principal eigenvectors ϕ_1 through ϕ_4 for 1000 samples for five choices of γ between 1 and 3 for algorithm (1) with η_k=1/(1000+k). Clearly, the convergence improves; i.e., the direction cosine goes to one faster as γ increases. Figure 2 shows the same results for algorithm 2.

Figure 1. Iterates of ϕ_1 through ϕ_4 for different γ for algorithm (1) with η_k=1/(1000+k). (see [2])

Figure 2. Iterates of ϕ_1 through ϕ_4 for different γ for algorithm (2) with η_k=1/(1000+k). (see [2])

Figure 3 shows the iterates of the first four principal eigenvectors ϕ_1 through ϕ_4 for 1000 samples for five choices of η_k for algorithm (1) with γ=1. Once again, the direction cosine goes to one faster for larger values of η_k up to an upper bound of η_k. Figure 4 shows the same results for algorithm (2).

Figure 3. Iterates of ϕ_1 through ϕ_4 for different η_k for algorithm (1) with γ=1. (see [2])

Figure 4. Iterates of ϕ_1 through ϕ_4 for different η_k for algorithm (2) with γ=1. (see [2])

The comparison of algorithms (1) and (2) for 1000 iterations for various γ and η_k can be obtained by comparing Figures 1 and 2 and also Figures 3 and 4. We see that algorithm (2) converges faster than algorithm (1) for the minor eigenvectors, corroborating our observations in Section 4.3.

References

[1] P.Baldi and K.Hornik, "Learning in Linear Neural Networks: A Survey", *IEEE Transactions on Neural Networks*, Vol. 6, No. 4, pp. 837-857, 1995.

[2] C.Chatterjee, V.P.Roychowdhury, E.K.P.Chong, "On Relative Convergence Properties of Principal Component Analysis Algorithms", submitted to *IEEE Transactions on Neural Networks*.

[3] V.Fabian, "On Asymptotic Normality in Stochastic Approximation", *The Annals of Mathematical Statistics*, Vol. 39, No. 4, pp. 1327-1332, 1968.

[4] G.H.Golub and C.F.VanLoan, *Matrix Computations*, Baltimore, MD: Johns Hopkins University Press, 1983.

[5] S.Haykin, *Neural Networks - A Comprehensive Foundation*, Maxwell Macmillan International, New York, 1994.

[6] E.Oja and J.Karhunen, "On Stochastic Approximation of the Eigenvectors and Eigenvalues of the Expectation of a Random Matrix", *Journ. of Math. Anal. Appl.*, Vol. 106, pp. 69-84, 1985.

[7] T.Okada and S.Tomita, "An Optimal Orthonormal System for Discriminant Analysis", *Pattern Recognition*, Vol. 18, No. 2, pp. 139-144, 1985.

[8] T.D.Sanger, "Optimal Unsupervised Learning in a Single-Layer Linear Feedforward Neural Network", *Neural Networks*, Vol. 2, pp. 459-473, 1989.

[9] L.Xu, "Least Mean Square Error Reconstruction Principle for Self-Organizing Neural-Nets", *Neural Networks*, Vol. 6, pp. 627-648, 1993.