Reprinted from JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS
Vol. 106, No. 1, February 15, 1985
Printed in Belgium

On Stochastic Approximation of the Eigenvectors and Eigenvalues of the Expectation of a Random Matrix

ERKKI OJA

Department of Applied Mathematics, University of Kuopio, P.O.B. 6, 70211 Kuopio 21, Finland

AND

JUHA KARHUNEN

Department of Technical Physics, Helsinki University of Technology, 02150 Espoo 15, Finland

Submitted by H. Kushner

In applications of signal processing and pattern recognition, eigenvectors and eigenvalues of the statistical mean of a random matrix sequence are needed. Iterative methods are suggested and analyzed, in which no sample moments are used. Convergence is shown by stochastic approximation theory. © 1985 Academic Press, Inc.

1. INTRODUCTION

There are several applications of digital signal processing and pattern recognition in which eigenvalues and eigenvectors of data correlation or covariance matrices are needed. Some such applications are optimal feature extraction in pattern recognition [2]; data compression and coding [19]; optimal pattern classification [8, 18]; antenna array processing for noise analysis and source location [14]; and adaptive spectral analysis for frequency estimation [15, 20, 16]. In a stationary case, the problem can be presented in the following general form: Consider an almost surely symmetric real $n \times n$ random matrix whose finite mean is denoted A. We want to compute the dominant eigenvalues and corresponding eigenvectors of A in a situation in which A itself is unknown but in which there is available a sequence of samples A_k, $k = 1, 2, \ldots$ with $E\{A_k\} = A$ for all k.

The straightforward method is to compute the sample mean and then use standard techniques like the QR method. This may be recommended if the $\{A_k\}$ sequence is completely general. However, in the applications

0022-247X/85 $3.00

involving correlation or covariance matrices, the A_k matrices have a specific form $A_k = u_k u_k^T$ with $\{u_k\}$ a random vector sequence. Then an iterative method which updates the estimates every time a new sample u_k becomes available has computational advantages [6].

As a stochastic approximation counterpart of the "simultaneous iteration method" of numerical analysis [17] we suggest the following algorithm:

$$\tilde{X}_k = X_{k-1} + A_k X_{k-1} \Gamma_k, \tag{1}$$

$$X_k = \tilde{X}_k R_k^{-1}, \tag{2}$$

in which $X_k = (x_k^{(1)} x_k^{(2)} \cdots x_k^{(s)}) \in \mathscr{R}^{n \times s}$ is a matrix whose columns $x_k^{(i)} \in \mathscr{R}^n$ are orthonormal and approximate s (with $s \leqslant n$) of the eigenvectors of A. In (2), R_k^{-1} is a matrix orthonormalizing the columns of $\tilde{X}_k$. Matrix $\Gamma_k \in \mathscr{R}^{s \times s}$ is the usual diagonal gain matrix of stochastic approximation.

In the present paper the almost sure convergence of the $x_k^{(i)}$ to eigenvectors of A is shown. These eigenvectors correspond to the s largest eigenvalues of A, which are assumed distinct, i.e. of unit multiplicity. It is also shown that the algorithm

$$\sigma_k^{(i)} = (1 - \gamma_k)\, \sigma_{k-1}^{(i)} + \gamma_k (x_{k-1}^{(i)T} A_k x_{k-1}^{(i)}) \qquad (i = 1, 2, \ldots, s) \tag{3}$$

then converges almost surely to the corresponding eigenvalues. The emphasis of this paper is on convergence theorems, with references to numerical applications.

The relation of (1), (2) to the simultaneous iteration method, which is an extension of the power method of numerical analysis, is of a theoretical nature only. There exist iterative methods well known in statistical literature, which use the power method directly for computing eigenvalues and eigenvectors of covariance matrices [1, 22]. These methods use a fixed sample of data vectors. The eigenvector and eigenvalue estimates are computed one at a time and their consistency follows from the consistency of sample moments. The algorithm given in the present paper is very different. No sample moments are computed, and several eigenvalues and eigenvectors are produced in a fully parallel manner.

Depending on the form of orthonormalization in (2), the present algorithm allows comparisons between some related stochastic approximation type algorithms reported earlier, as well as between the asymptotic solutions $x_k^{(i)}$ and the asymptotic paths of ordinary differential equations. Krasulina [8] introduced a stochastic approximation algorithm for computing one dominant eigenvalue and the corresponding eigenvector of A:

$$x_k = x_{k-1} + \gamma_k \left[A_k x_{k-1} - \frac{x_{k-1}^T A_k x_{k-1}}{x_{k-1}^T x_{k-1}} x_{k-1} \right], \tag{4}$$

where $\gamma_k \geqslant 0$ is a sequence of gain scalars. The convergence of x_k to a random vector lying in the eigenspace corresponding to the largest eigenvalue of $E\{A_k\}$ follows from the inequality

$$E\{\|x_{k+1}\|^2 | x_k\} \leqslant \|x_k\|^2(1+\gamma_{k+1}^2 E\{\|A_{k+1}\|^2\}). \tag{5}$$

If $E\{\|A_k\|^2\}$ is bounded and $\sum \gamma_k^2$ converges, this yields convergence, but the upper limit for $E\{\|x_k\|^2\}$ can be very large. Computer simulations confirm this.

We discuss in Section 5 of the present paper a simpler algorithm

$$x_k = x_{k-1} + \gamma_k[A_k x_{k-1} - (x_{k-1}^T A_k x_{k-1})\, x_{k-1}], \tag{6}$$

whose convergence to a *unit* eigenvector of A emerges as a corollary of results established in Section 2.

Algorithm (1), (2) is also closely related to a data orthogonalization method given by Owsley [14] in context of signal processing. His algorithm is a special case of (1), (2) with $A_k = u_k u_k^T$, all diagonal elements of Γ_k equal and constant, and R_k^{-1} performing Gram–Schmidt orthonormalization. Also, Thompson [20] gives essentially the same algorithm with $A_k = -u_k u_k^T$, although vector u_k then has different properties. Geometrical considerations have been presented by both Owsley and Larimore and Calvert [10]. However, the authors do not give a rigorous proof of convergence of algorithm (1), (2).

Our method of proof relies on results given by Kushner and Clark [9], concerning almost sure convergence of stochastic approximation algorithms. We prefer this technique to the classical methods mostly based on Dvoretzky's results (see, e.g., [21]), because the use of limiting differential equations seems to provide a much better insight into the asymptotic behavior and mutual relations of the algorithms under study.

2. Convergence of the Unit Eigenvector Corresponding to the Largest Eigenvalue

When X_k consists of one column x_k only, Eqs. (1), (2) read

$$\tilde{x}_k = x_{k-1} + \gamma_k A_k x_{k-1}, \tag{7}$$

$$x_k = \tilde{x}_k / \|\tilde{x}_k\|, \tag{8}$$

where the Euclidean vector norm is used. Assuming γ_k small enough, (7) and (8) can be expanded as a power series in γ_k, yielding

$$x_k = x_{k-1} + \gamma_k[A_k x_{k-1} - (x_{k-1}^T A_k x_{k-1})\, x_{k-1}] + \gamma_k b_k. \tag{9}$$

There $b_k = 0(\gamma_k)$. Since $x_{k-1}^T x_{k-1} = 1$, Eq. (9) can further be written as

$$x_k = x_{k+1} + \gamma_k \left[Ax_{k-1} - \frac{(x_{k-1}^T A x_{k-1})}{x_{k-1}^T x_{k-1}} x_{k-1} \right] + \gamma_k [(A_k - A) x_{k-1} - x_{k-1}^T (A_k - A) x_{k-1} x_{k-1}] + \gamma_k b_k. \quad (10)$$

Assume now:

A1. *Each A_k is almost surely bounded and symmetric and the A_k are mutually statistically independent with $E\{A_k\} = A$ for all k.*

A2. *The largest eigenvalue of A has unit multiplicity.*

A3. $\gamma_k \geq 0, \sum \gamma_k^2 < \infty, \sum \gamma_k = \infty$.

A4. *Each A_k has a probability density which is bounded away from zero uniformly in k in some neighbourhood of A in $\mathcal{R}^{n \times n}$.*

We modify a result given by Kushner and Clark [9, p. 39] to suit the present algorithm:

LEMMA 1. *Let* A1 *and* A3 *hold. Let z_0 be a locally asymptotically stable (in the sense of Liapunov) solution to*

$$\frac{dz}{dt} = Az - \frac{(z^T A z) z}{z^T z} \quad (11)$$

with domain of attraction $\mathcal{D}(z_0)$. If there is a compact set $\mathcal{A} \subset \mathcal{D}(z_0)$ such that the solution of (7), (8) *satisfies $P\{x_k \in \mathcal{A}$ infinitely often$\} = 1$, then x_k tends to z_0 almost surely.*

Proof. The boundedness of x_k is trivially true due to (8). Assumptions A.2.2.1 and A.2.2.3 of Theorem 2.3.1 in Kushner and Clark [9] follow directly from (11) and A3. Condition A.2.2.2 is verified as follows: in (9), we have

$$b_k = -1/2\gamma_k (x_{k-1}^T A_k^2 x_{k-1}) x_{k-1} - 1/2\gamma_k \beta_k A_k x_{k-1} + \gamma_k^{-1} [(1 + \gamma_k \beta_k)^{-1/2} - 1 + 1/2\gamma_k \beta_k](I - \gamma_k A_k) x_{k-1}$$

with $\beta_k = 2x_{k-1}^T A_k x_{k-1} + \gamma_k x_{k-1}^T A_k^2 x_{k-1}$. Since x_{k-1} and A_k are a.s. bounded, b_k is a.s. bounded and tends to zero as $\gamma_k \to 0$. Condition A.2.2.4 is finally verified as follows:

$$\sum_{i=k}^{m} \gamma_i [(A_i - A) x_{i-1} - x_{i-1}^T (A_i - A) x_{i-1} x_{i-1}]$$

is a martingale sequence due to the independence and a.s. boundedness of

matrices A_k and, as pointed out by Kushner and Clark, for any $\varepsilon > 0$ it holds that

$$\lim_{k \to \infty} P\left\{ \sup_{m \geqslant k} \left\| \sum_{i=k}^{m} \gamma_i [(A_i - A)\, x_{i-1} - x_{i-1}^T (A_i - A)\, x_{i-1} x_{i-1}] \right\| \geqslant \varepsilon \right\} = 0$$

because $\sum \gamma_k^2$ converges ([3] or [11]).

We next show that the unit eigenvectors of A corresponding to the largest eigenvalue are indeed the possible limits of the O.D.E. in Lemma 1.

LEMMA 2. *In the O.D.E.* (11), *let* A2 *hold and let* $c^{(1)}$ *be one of the two unit eigenvectors corresponding to the largest eigenvalue* $\lambda^{(1)}$ *of matrix* A. *The points* $c^{(1)}$ *and* $-c^{(1)}$ *are (uniformly) asymptotically stable. The domain of attraction of* $c^{(1)}$ *is* $\mathscr{D}(c^{(1)}) = \{x \in \mathscr{R}^n \mid x^T c^{(1)} > 0\}$ *and that of* $-c^{(1)}$ *is* $\mathscr{D}(-c^{(1)}) = \{x \in \mathscr{R}^n \mid x^T c^{(1)} < 0\}$.

Proof. Set

$$z(t) = \sum_{i=1}^{n} \eta^{(i)}(t)\, c^{(i)}$$

with $c^{(1)}, \ldots, c^{(n)}$ an orthonormal set of eigenvectors of A. Then (11) yields $d\eta^{(i)}/dt = \lambda^{(i)} \eta^{(i)} - (z^T A z)\, \eta^{(i)}/z^T z$, $i = 1, \ldots, n$. The solutions $\eta^{(i)}(t)$ are unique, and if $\eta^{(i)}(t_0) = 0$ for some t_0, then $\eta^{(i)}(t)$ is identically zero. For simplicity, set $t_0 = 0$. If now $z(0)^T c^{(1)} = \eta^{(1)}(0) = 0$, then $\eta^{(1)}(t)$ remains zero for all t and $z(t)$ cannot tend to $c^{(1)}$ or $-c^{(1)}$. Assume now that $\eta^{(1)}(0) \neq 0$. Then $\eta^{(1)}(t) \neq 0$ for all t and we may define $\zeta^{(i)}(t) = \eta^{(i)}(t)/\eta^{(1)}(t)$, yielding $d\zeta^{(i)}(t)/dt = (d\eta^{(i)}/dt\, \eta^{(1)} - \eta^{(i)}\, d\eta^{(1)}/dt)/\eta^{(1)2}$, hence

$$d\zeta^{(i)}/dt = (\lambda^{(i)} - \lambda^{(1)})\, \zeta^{(i)} \tag{12}$$

whose solution on $[0, \infty)$ is

$$\zeta^{(i)}(t) = \exp[(\lambda^{(i)} - \lambda^{(1)})\, t]\, \zeta^{(i)}(0). \tag{13}$$

There $\lambda^{(i)}$ is the eigenvalue of A corresponding to $c^{(i)}$. Because $\lambda^{(i)} < \lambda^{(1)}$, $\zeta^{(i)}(t)$ tends to zero as $t \to \infty$ for all $i = 2, \ldots, n$. On the other hand, (11) implies $(d/dt)\, \|z\|^2 = 2z^T (dz/dt) = 2(z^T A z - z^T A z) = 0$. Thus if $\|z(0)\| = 1$, then $\|z(t)\| = 1$ for all t. Then $\sum_{i=1}^{n} \eta^{(i)}(t)^2 = 1$, hence the convergence of $\zeta^{(i)}(t)$ to zero $(i = 2, \ldots, n)$ implies the convergence of $\eta^{(i)}(t)$ to zero $(i = 2, \ldots, n)$ as $t \to \infty$. But then $\lim_{t \to \infty} \eta^{(1)}(t)^2 = 1$. Since $\eta^{(1)}(t) \neq 0$ for all t, we have $\lim \eta^{(1)}(t) = \pm 1$ according to the sign of $\eta^{(1)}(0) = z(0)^T c^{(1)}$. This concludes the proof.

LEMMA 3. *In* (7), (8), *let* A1 *to* A4 *hold. Then there exists a number* ε *such that the event* $|x_k^T c^{(1)}| \geqslant \varepsilon$ *occurs infinitely often almost surely.*

Proof. Equations (7) and (8) yield

$$x_k^T c^{(1)} = \frac{x_{k-1}^T c^{(1)} + \gamma_k c^{(1)T}[(A_k - A) + A]\, x_{k-1}}{(1 + \gamma_k A_k)\, x_{k-1}}. \tag{14}$$

By assumption A4, there exist positive numbers δ and ρ such that $P\{c^{(1)T}(A_k - A)\, x_{k-1} \geq \delta\} \geq \rho$, uniformly in k. Assume without loss of generality that $x_{k-1}^T c^{(1)} > 0$. Let α be an almost sure upper bound for $\|A_k\|$ and denote again the largest eigenvalue of A by $\lambda^{(1)}$. Then we obtain from Eq. (14)

$$x_k^T c^{(1)} \geq \frac{1}{1 + \gamma_k \alpha} (c^{(1)T} x_{k-1} + \gamma_k \lambda^{(1)} c^{(1)T} x_{k-1} + \gamma_k \delta)$$

$$= \frac{1 + \gamma_k \lambda^{(1)}}{1 + \gamma_k \alpha} c^{(1)T} x_{k-1} + \frac{\gamma_k}{1 + \gamma_k \alpha} \delta.$$

Since matrices A_j are statistically independent, with probability at least equal to ρ^{M-k+1} we have $c^{(1)T}(A_j - A)\, x_{j-1} \geq \delta$ for all $j = k, k+1, \ldots, M$. Then

$$x_j^T c^{(1)} \geq \frac{1 + \gamma_j \lambda^{(1)}}{1 + \gamma_j \alpha} c^{(1)T} x_{j-1} + \frac{\gamma_j}{1 + \gamma_j \alpha} \delta \qquad \text{for} \quad j = k, \ldots, M,$$

implying that

$$x_M^T c^{(1)} \geq \prod_{j=k}^{M} \left(\frac{1 + \gamma_j \lambda^{(1)}}{1 + \gamma_j \alpha} \right) c^{(1)T} x_{k-1}$$

$$+ \delta \sum_{j=k}^{M} \left(\frac{\gamma_j}{1 + \gamma_j \alpha} \right) \prod_{i=j+1}^{M} \left(\frac{1 + \gamma_i \lambda^{(1)}}{1 + \gamma_i \alpha} \right)$$

$$\geq \delta \sum_{j=k}^{M} \left(\frac{\gamma_j}{1 + \gamma_j \alpha} \right) \prod_{i=j+1}^{M} \left(\frac{1 + \gamma_i \lambda^{(1)}}{1 + \gamma_i \alpha} \right),$$

since $c^{(1)T} x_{k-1}$ was assumed positive and due to A3 it may be assumed without loss of generality that $0 \leq \gamma_i \leq |1/\lambda^{(1)}|$ for $i \geq k$. In the above, we define a product of the form $\prod_{i=M+1}^{M}$ to have the value 1, as usual. Furthermore,

$$\delta \sum_{j=k}^{M} \left(\frac{\gamma_j}{1 + \gamma_j \alpha} \right) \prod_{i=j+1}^{M} \left(\frac{1 + \gamma_i \lambda^{(1)}}{1 + \gamma_i \alpha} \right)$$

$$= \frac{\delta}{\alpha - \lambda^{(1)}} \sum_{j=k}^{M} \frac{(1 + \gamma_j \alpha) - (1 + \gamma_j \lambda^{(1)})}{1 + \gamma_j \alpha} \prod_{i=j+1}^{M} \left(\frac{1 + \gamma_i \lambda^{(1)}}{1 + \gamma_i \alpha} \right)$$

$$= \frac{\delta}{\alpha - \lambda^{(1)}} \left(1 - \prod_{i=k}^{M} \left(\frac{1 + \gamma_i \lambda^{(1)}}{1 + \gamma_i \alpha} \right) \right).$$

Since α is now an upper bound for A_k it follows that α can be chosen larger than $|\lambda^{(1)}|$. Then there exists a number $\theta > 0$ such that $e^{-\theta\xi} \geqslant (1+\lambda^{(1)}\xi)/(1+\alpha\xi)$ for all ξ in the interval $[0, |1/\lambda^{(1)}|]$. This implies

$$\frac{1+\gamma_i\lambda^{(1)}}{1+\gamma_i\alpha} \leqslant e^{-\theta\gamma_i}$$

for $i \geqslant k$ and

$$\prod_{i=k}^{M}\left(\frac{1+\gamma_i\lambda^{(1)}}{1+\gamma_i\alpha}\right) \leqslant \exp\left(-\theta\sum_{i=k}^{M}\gamma_i\right),$$

implying

$$c^{(1)T}x_M \geqslant \frac{\delta}{\alpha-\lambda^{(1)}}\left(1-\exp\left(-\theta\sum_{i=k}^{M}\gamma_i\right)\right).$$

Choose now $\varepsilon = \frac{1}{2}\delta/(\alpha-\lambda^{(1)})$. Due to the divergence of the sum $\sum\gamma_i$, there is an index $\bar{M}$ such that

$$\frac{\delta}{\alpha-\lambda^{(1)}}\left(1-\exp\left(-\theta\sum_{i=k}^{\bar{M}}\gamma_i\right)\right) \geqslant \varepsilon.$$

The conclusion from the above is that the event

$$\{c^{(1)T}x_{\bar{M}} \geqslant \varepsilon, \text{ when } c^{(1)T}x_{k-1} > 0\},$$

with ε a fixed positive number, has at least probability $\rho^{\bar{M}-k+1}$. Since $\{x_k\}$ is a Markov process due to the statistical independence of the A_k, it follows that starting from any state x_{k-1} such that $c^{(1)T}x_{k-1} > 0$, the region $\{x \mid x^Tc^{(1)} \geqslant \varepsilon\}$ is eventually reached with probability one [3]. The proof is completely analogous for the case $c^{(1)T}x_{k-1} < 0$ and the region $\{x \mid x^Tc^{(1)} \leqslant -\varepsilon\}$. So the union of these two regions is reached by the process $\{x_k\}$ infinitely often with probability one, as was to be shown.

The convergence of the algorithm (7), (8) is now a direct corollary of the above lemmas.

THEOREM 1. *In algorithm* (7), (8), *let* A1, A2, A3, *and* A4 *hold. Then* x_k *tends either to* $c^{(1)}$ *or* $-c^{(1)}$ *almost surely as* $k \to \infty$.

Proof. By Lemma 3, $\{x_k\}$ visits a.s. infinitely often a compact subset of the domain of attraction of one of the asymptotically stable points $c^{(1)}$ and $-c^{(1)}$ in differential equation (11). Lemma 1 implies then the theorem.

3. Determination of all Eigenvectors

In establishing convergence for the second, third, etc. eigenvector, we proceed along very similar lines as in the case of the first vector. Assumptions A2 and A3 must first be modified to suit the algorithm (1), (2). They are now replaced by

A5. *The s largest eigenvalues of A are positive and each of unit multiplicity.*

A6. *The diagonal elements of the $s \times s$ diagonal matrix Γ_k are, in this order, $\gamma_k, \theta^{(2)}\gamma_k, \ldots, \theta^{(s)}\gamma_k$, with each $\theta^{(i)}$ positive and γ_k satisfying* A3.

Lemma 3. *For γ_k small, the jth column of X_k in* (1), (2) *satisfies*

$$x_k^{(j)} = x_{k-1}^{(j)} + \theta^{(j)}\gamma_k [I - x_{k-1}^{(j)} x_{k-1}^{(j)T} - \sum_{i=1}^{j-1} (1 + \theta^{(i)}/\theta^{(j)})\, x_{k-1}^{(i)} x_{k-1}^{(i)T}]\, A_k x_{k-1}^{(j)} + O(\gamma_k^2). \tag{15}$$

Proof. It is easily shown that (15) holds for $j=1$ (this is then Eq. (9)). Equation (15) for $j=2,\ldots, s$ can be shown by induction, making use of the orthonormality of vectors $x_{k-1}^{(i)}$ and $x_{k-1}^{(j)}$ for $i<j$.

In exactly the same way as Eq. (11) is derived from (9), Eq. (15) corresponds to the O.D.E.

$$dz^{(j)}/dt = \theta^{(j)}[Az^{(j)} - (z^{(j)T}Az^{(j)})\, z^{(j)} - \sum_{i=1}^{j-1} (1 + \theta^{(i)}/\theta^{(j)})(z^{(i)T}Az^{(j)})\, z^{(i)}] \tag{16}$$

whose asymptotically stable solutions are the possible almost sure limits for $x_k^{(j)}$ as k grows to infinity. These stable points are given in the following.

Lemma 4. *In the set of differential equations* (16) *for $j=1,\ldots, s$, let* A5 *and* A6 *hold. Let $c^{(1)},\ldots, c^{(s)}$ be unit eigenvectors corresponding to the s largest eigenvalues of A. In the jth equation, the points $c^{(j)}$ and $-c^{(j)}$ are asymptotically stable.*

Proof. Denote $e^{(j)}(t) = z^{(j)}(t) - c^{(j)}$. Let $\lambda^{(j)}$ denote the eigenvalue corresponding to $c^{(j)}$. We have from (16)

$$de^{(j)}/dt = B^{(j)}e^{(j)} + \sum_{i=1}^{j-1} C^{(i,j)}e^{(i)} + f^{(j)}(e^{(1)},\ldots, e^{(s)}) \tag{17}$$

with

$$B^{(j)} = \theta^{(j)}[A - \lambda^{(j)} I - 2\lambda^{(j)} c^{(j)} c^{(j)T} - \sum_{i=1}^{j-1} (1 + \theta^{(i)}/\theta^{(j)})\, \lambda^{(i)} c^{(i)} c^{(i)T}], \tag{18}$$

$$C^{(i,j)} = -(\theta^{(j)} + \theta^{(i)})\, \lambda^{(j)} c^{(i)} c^{(j)T}, \tag{19}$$

and

$$\begin{aligned} f^{(j)} = \theta^{(j)}[&-e^{(j)}(e^{(j)T} A e^{(j)}) - 2\lambda^{(j)}(e^{(j)T} c^{(j)})\, e^{(j)} \\ &- (e^{(j)T} A e^{(j)})\, c^{(j)}] - \sum_{i=1}^{j-1} (\theta^{(j)} + \theta^{(i)}) \\ &\times [(e^{(i)T} A e^{(j)})\, e^{(i)} + \lambda^{(j)}(e^{(i)T} c^{(j)})\, e^{(i)} \\ &+ \lambda^{(i)}(e^{(j)T} c^{(i)})\, e^{(i)} + (e^{(i)T} A e^{(j)})\, c^{(i)}]. \end{aligned} \tag{20}$$

If we denote $e^T = (e^{(1)T} e^{(2)T} \cdots e^{(s)T})$, defining e as a ns-dimensional vector function, we have

$$de/dt = De + f(e) \tag{21}$$

with $D \in \mathscr{R}^{ns \times ns}$ a matrix of lower triangular block form whose diagonal blocks are the matrices $B^{(1)},..., B^{(s)}$ and $f(e) \in \mathscr{R}^{ns}$ a vector with the $f^{(1)},..., f^{(s)}$ as its subvectors. Now both $f(e)$ and $\lim \partial f/\partial e$ are zero at $e = 0$, due to (20). The eigenvalues of D are the eigenvalues of the diagonal blocks $B^{(1)}$ to $B^{(s)}$. Each of these matrices has the same vectors $c^{(1)},..., c^{(n)}$ as eigenvectors, as is apparent from (18). The eigenvalue of $B^{(j)}$ corresponding to eigenvector $c^{(i)}$ equals $-\theta^{(i)}\lambda^{(i)} - \theta^{(j)}\lambda^{(j)}$ for $i < j$, $-2\theta^{(j)}\lambda^{(j)}$ for $i = j$, and $\theta^{(j)}(\lambda^{(i)} - \lambda^{(j)})$ for $i > j$. Due to A5, A6, all of these (for $j \leqslant s, i \leqslant s$) are negative. The asymptotic stability of zero as the solution of (21) follows from Theorem 2.4 of Hale [4, p. 86]. This concludes the proof.

Referring again to Theorem 2.3.1 of Kushner and Clark [9], the convergence of algorithm (1), (2) may be established.

THEOREM 2. *Assume* A1, A3, A5, *and* A6 *in algorithm* (1), (2), *and assume that with probability one each process* $\{x_k^{(j)}\}$ $(j = 1,..., s)$ *visits infinitely often a compact subset of the domain of attraction of one of the asymptotically stable points, say* $+c^{(j)}$. *Then almost surely*

$$\lim_{k \to \infty} x_k^{(j)} = c^{(j)} \qquad (j = 1,..., s). \tag{22}$$

Remark. It is immaterial in view of applications whether the limit is $c^{(j)}$

or $-c^{(j)}$. The assumption of $x_k^{(j)}$ coming infinitely often close enough to its eventual limit is in fact an assumption on the distributions of the $\{A_k\}$ sequence. Since $c^{(j)}$ is an eigenvector of A corresponding to a strictly positive eigenvalue, and hence $E\{c^{(j)T}A_k c^{(j)}\}$ is positive, $c^{(j)T}A_k c^{(j)}$ must be "large" infinitely often. The increment in algorithm (1) then tends to bring $x_k^{(j)}$ closer and closer to either $c^{(j)}$ or $-c^{(j)}$. In computer simulations, no problems related to this assumption ever occur. The validity of this assumption in algorithm (7), (8) for computing one eigenvector was shown above in Lemma 3 under assumption A4.

4. Determination of the Eigenvalues

Next we turn to algorithm (3). We have

THEOREM 3. *Let* A1 *and* A3 *hold, and assume in* (3) *that* $x_k^{(i)}$, *given by algorithm* (1), (2), *tends almost surely to an eigenvector of* A *corresponding to eigenvalue* $\lambda^{(i)}$. *Let* $\sigma_0^{(i)}$ *be a.s. bounded. Then* $\sigma_k^{(i)}$ *is a.s. uniformly bounded and almost surely*

$$\lim_{k\to\infty} \sigma_k^{(i)} = \lambda^{(i)}.$$

Proof. For convenience, in the following proof the superscript i has been dropped, since each $\sigma_k^{(i)}$ $(i=1,\ldots,s)$ may be considered separately. Equation (3) yields

$$\sigma_k = \prod_{j=1}^{k}(1-\gamma_j)\,\sigma_0 + \sum_{j=1}^{k}\gamma_j(x_{j-1}^T A_j x_{j-1}) \prod_{h=j+1}^{k}(1-\gamma_h),$$

hence almost surely

$$|\sigma_k| \leqslant \left|\prod_{j=1}^{k}(1-\gamma_j)\right| |\sigma_0| + \alpha \sum_{j=1}^{k}\gamma_j \left|\prod_{h=j+1}^{k}(1-\gamma_h)\right|,$$

where α is the a.s. upper bound of $\|A_k\|$. Note that $\|x_{j-1}\| = 1$. Assumption A3 implies that from some index K, we can assume $0 \leqslant \gamma_k \leqslant 1$; since then

$$\prod_{j=K}^{k}(1-\gamma_j) \leqslant 1, \qquad \sum_{j=K}^{k}\gamma_j \prod_{h=j+1}^{k}(1-\gamma_h) = 1 - \prod_{j=K}^{k}(1-\gamma_j) \leqslant 1,$$

we obtain the almost sure bound

$$|\sigma_k| \leqslant |\sigma_K| + \alpha,$$

showing the first part of the theorem. To establish almost sure convergence, write (3) in the form

$$\begin{aligned}\sigma_k &= \sigma_{k-1} - \gamma_k[\sigma_{k-1} - \lambda + (\lambda - x_{k-1}^T A x_{k-1}) + x_{k-1}^T (A - A_k)\, x_{k-1}] \\ &= \sigma_{k-1} - \gamma_k[\sigma_{k-1} - \lambda + \nu_k + \xi_k].\end{aligned}$$

Due to the a.s. convergence of x_{k-1} to c, $\nu_k = \lambda - x_{k-1}^T A x_{k-1}$ tends to zero a.s. Let $\mathcal{S}_{k-1}$ be the σ-algebra generated by $A_1, \ldots, A_{k-1}$. Then all x_{k-1}, $x_{k-2}, \ldots$ are $\mathcal{S}_{k-1}$-measurable due to A1, and $\xi_k = x_{k-1}^T (A - A_k)\, x_{k-1}$ satisfies $E\{\xi_k \mid \xi_{k-1}, \xi_{k-2}, \ldots\} = E\{\xi_k \mid \mathcal{S}_{k-1}\} = 0$. Also, because $\|x_{k-1}\| = 1$ and A_k is a.s. bounded, each ξ_k has bounded variance. Then $\sum \gamma_k \xi_k$ is a martingale sequence and we have

$$\lim_{p \to \infty} P\left\{\sup_{q \geq p} \left| \sum_{k=p}^{q} \gamma_k \xi_k \right| \geq \varepsilon \right\} = 0$$

for all $\varepsilon > 0$, since $\sum \gamma_k^2$ converges. Algorithm (3) thus satisfies A.2.2.1 through A.2.2.4 of Theorem 2.3.1 of Kushner and Clark [9]. Since the only asymptotically stable solution of the O.D.E.

$$d\sigma/dt = \lambda - \sigma \tag{23}$$

is λ, whose domain of attraction is the whole real line, the a.s. convergence of σ_k to λ has been established.

5. Some Modifications of the Basic Algorithm

Another similar recursive method to compute the eigenvector corresponding to the largest eigenvalue is suggested by the asymptotic analysis of Sec. 2 and given by Eq. (9). When the $O(\gamma_k^2)$ term is dropped there we have

$$\begin{aligned}x_k &= x_{k-1} + \gamma_k[A_k x_{k-1} - x_{k-1}^T A_k x_{k-1} x_{k-1}] \\ &= x_{k-1} + \gamma_k[A x_{k-1} - x_{k-1}^T A x_{k-1} x_{k-1}] \\ &\quad + \gamma_k[(A_k - A)\, x_{k-1} - x_{k-1}^T (A_k - A)\, x_{k-1} x_{k-1}].\end{aligned} \tag{24}$$

This shows that the limiting differential equation is now

$$\frac{dz}{dt} = Az - (z^T A z)\, z.$$

Its asymptotically stable points are again $c^{(1)}$ and $-c^{(1)}$ with the same domains of attraction as in Eq. (11), as can be shown in analogy with Lemma 2. To show convergence to $c^{(1)}$ in (24) we have to verify that x_k remains bounded; the rest of the proof goes through as before with minor variations.

In (7), (8) the boundedness was guaranteed by an explicit normalization at each step. No such normalization is present in (24). It turns out that, even with γ_k bounded, there is a possibility that during the early phase of the recursion $\|x_k\|$ grows too large to be able to catch up any more with the orbit of the limiting O.D.E. This must be prevented by setting a specific upper bound for γ_k. Also, there is a possibility that $\|x_k\|$ grows even then unless A_k is positive semidefinite, although in practice this does not seem to be a necessary assumption. We show the following:

LEMMA 5. *In* (24), *let* A_k *be positive semidefinite and bounded almost surely for all* k. *Let* $\gamma_k \geqslant 0$. *Assume that* x_0 *is a.s. bounded. Then there exists a uniform upper bound for* γ_k *such that* x_k *is a.s. uniformly bounded.*

Proof. Let μ be a real number satisfying $\|x_0\| - 1 \leqslant \mu$ and $\mu^3 - \mu^2 \geqslant 8$. Let α be the a.s. upper bound for $\|A_k\|$. We will show by a simple induction argument that $\|x_k\|^2 \leqslant \mu + 1$ if

$$\gamma_k \leqslant \frac{2}{\alpha\mu}. \tag{25}$$

Equation (24) yields

$$\begin{aligned}\|x_k\|^2 = \|x_{k-1}\|^2 &+ 2\gamma_k(1 - \|x_{k-1}\|^2)(x_{k-1}^T A_k x_{k-1}) \\ &+ \gamma_k^2 x_{k-1}^T A_k^2 x_{k-1} + \gamma_k^2(\|x_{k-1}\|^2 - 2)(x_{k-1}^T A_k x_{k-1})^2. \end{aligned} \tag{26}$$

First, assume that $\|x_{k-1}\| < 1$. Then we have a.s.

$$\begin{aligned}\|x_k\|^2 &\leqslant 1 + 2\gamma_k \|A_k\| + \gamma_k^2 \|A_k\|^2 \leqslant 1 + 2\gamma_k\alpha + \gamma_k^2\alpha^2 \\ &\leqslant 1 + 4/\mu + 4/\mu^2 \leqslant 2(1 + 4/\mu^2) \leqslant \mu + 1\end{aligned}$$

because $\mu^3 - \mu^2 \geqslant 8$. Second, assume $1 \leqslant \|x_{k-1}\|^2 < 2$. Then a.s.

$$\|x_k\|^2 \leqslant 2 + 2\gamma_k^2\alpha^2 \leqslant 2 + 8/\mu^2 \leqslant \mu + 1.$$

Finally, assume $2 \leqslant \|x_{k-1}\|^2 \leqslant \mu + 1$. From (26), $\|x_k\|^2 \leqslant \|x_{k-1}\|^2$ if

$$\begin{aligned}&2\gamma_k(1 - \|x_{k-1}\|^2)(x_{k-1}^T A_k x_{k-1}) + \gamma_k^2 x_{k-1}^T A_k^2 x_{k-1} \\ &\quad + \gamma_k^2(\|x_{k-1}\|^2 - 2)(x_{k-1}^T A_k x_{k-1})^2 \leqslant 0.\end{aligned}$$

Since $\gamma_k \geqslant 0$, a sufficient condition for the above is

$$\gamma_k \leqslant \frac{2(\|x_{k-1}\|^2 - 1)(x_{k-1}^T A_k x_{k-1})}{x_{k-1}^T A_k^2 x_{k-1} + (\|x_{k-1}\|^2 - 2)(x_{k-1}^T A_k x_{k-1})^2}. \tag{27}$$

The denumerator of (27) is bounded from above by

$$\begin{aligned} &x_{k-1}^T A_k^2 x_{k-1} + (\|x_{k-1}\|^2 - 2)(x_{k-1}^T A_k^2 x_{k-1})\,\|x_{k-1}\|^2 \\ &\quad = (1 - \|x_{k-1}\|^2)^2\,(x_{k-1}^T A_k^2 x_{k-1}), \end{aligned}$$

where the Cauchy–Schwartz inequality has been applied, and we have a.s.

$$\begin{aligned} &\frac{2(\|x_{k-1}\|^2 - 1)(x_{k-1}^T A_k x_{k-1})}{x_{k-1}^T A_k^2 x_{k-1} + (\|x_{k-1}\|^2 - 2)(x_{k-1}^T A_k x_{k-1})^2} \\ &\quad \geqslant \frac{2(x_{k-1}^T A_k x_{k-1})}{(x_{k-1}^T A_k^2 x_{k-1})(\|x_{k-1}\|^2 - 1)} \geqslant \frac{2}{\mu \lambda_{\max}(A_k)} \geqslant \frac{2}{\mu\alpha}. \end{aligned}$$

Since $\gamma_k \leqslant 2\mu^{-1}\alpha^{-1}$, Eq. (27) holds and $\|x_k\|^2 \leqslant \mu + 1$. Lemma 4 follows by induction.

The convergence of algorithm (24) is now a corollary of the results established in Section 2.

THEOREM 4. *In algorithm* (24), *let* A1, A2, *and* A3 *hold. Assume that each* A_k *is a.s. positive semidefinite,* γ_k *satisfies* (25) *and* $\|x_0\|^2$ *is a.s. bounded by* $\mu + 1$. *Assume further that for some positive* ε, *the event* $(x_k^T c^{(1)})^2 \geqslant \varepsilon$ *occurs infinitely often with probability one. Then* x_k *tends either to* $c^{(1)}$ *or to* $-c^{(1)}$ *almost surely as* $k \to \infty$.

Proof. When some obvious modifications are made in Lemma 1 and its proof, the above theorem follows from Lemmas 1 and 2 in the same way as Theorem 1.

For the other eigenvectors of A, two possible modifications of (1), (2) are the following:

$$\begin{aligned} x_k^{(j)} = x_{k-1}^{(j)} + \theta^{(j)}\gamma_k \Bigg[&A_k x_{k-1}^{(j)} - (x_{k-1}^{(j)T} A_k x_{k-1}^{(j)})\, x_{k-1}^{(j)} \\ &- \sum_{i=1}^{j-1} (1 + \theta^{(i)}/\theta^{(j)})(x_{k-1}^{(i)T} A_k x_{k-1}^{(j)})\, x_{k-1}^{(i)} \Bigg], \end{aligned} \tag{28}$$

which is simply Eq. (15) when the $O(\gamma_k^2)$ term has been dropped; and

$$X_k = X_{k-1} + \gamma_k [A_k X_k - (X_{k-1}^T A_k X_{k-1})\, X_{k-1}] \tag{29}$$

with

$$X_k = (x_k^{(1)} \cdots x_k^{(s)}).$$

Algorithm (29) does not produce the eigenvectors as such, but only an orthonormal basis of the subspace spanned by the eigenvectors $c^{(1)}$ to $c^{(s)}$. This may be sufficient in some applications, notably in the learning subspace methods of classification [7, 12], but it is not sufficient if the eigenvectors are needed.

6. Some Numerical Results on the Rate of Convergence and Estimation Errors

Algorithm (1), (2) has been used by Owsley [14] in a acoustic source location problem and by Thompson [20] in an adaptive implementation on Pisarenko's harmonic retrieval method to find the eigenvectors corresponding either to the largest or the smallest eigenvalues of a data correlation matrix. Results are given in the two papers referred to above. The smallest eigenvalue problem is converted to the largest eigenvalue problem when matrix $-A_k$ is used in Eq. (1). Both authors use constant gains γ.

In a computer test with artificial data, we used 15-component independent stationary sample vectors u_k to define matrices $A_k = u_k u_k^T \in \mathscr{R}^{15\times 15}$. Due to the form of (1), these matrices need not be formed explicitly. The largest eigenvalues of the theoretical correlation matrix $A = E\{u_k u_k^T\}$ were $\lambda^{(1)} = 2.613$ and $\lambda^{(2)} = 1.470$. With the gain sequence $\gamma_k^{(1)} = \gamma_k = 0.5/k$, the first eigenvector estimate $x_k^{(1)}$ converged as shown in Table I. The initial value $x_0^{(1)}$ was one of the sample vectors u_k. The convergence is fast in the beginning but then slows down. The gain of the form $1/k$ seems to be near optimal in practice.

TABLE I

Convergence Rate of Algorithm (1), (2)

k	$\|c^{(1)} - x_k^{(1)}\|$
30	0.2250
75	0.0991
150	0.0895
300	0.0884

For details, see text.

The standard method to estimate $c^{(1)}$ would be to first compute $\hat{A} = (1/K) \sum_{k=1}^{K} u_k u_k^T$ and then obtain its largest eigenvalue and eigenvector, e.g., by the power method. A comparison showed that the stochastic gradient algorithm needs roughly 1.5 to 2 times more sample vectors u_k to achieve the same estimation error, as compared to the standard method. However, this is compensated by the larger speed of computation. Also the storage demands are much smaller for algorithm (1), (2).

Results on the computation of several eigenvectors and also eigenvalues, both for stationary and nonstationary data and also using algorithms (24) or (28), have been given elsewhere by the present authors [5, 6].

REFERENCES

1. T. W. ANDERSON, "An Introduction to Multivariate Statistical Analysis," Wiley, New York, 1958.
2. Y. T. CHIEN AND K. S. FU, On the generalized Karhunen–Loève expansion, *IEEE Trans. Inform. Theory* **IT-13** (1967), 518–520.
3. J. L. DOOB, "Stochastic Processes," Wiley, New York, 1953.
4. J. K. HALE, "Ordinary Differential Equations," Wiley, New York, 1969.
5. J. KARHUNEN AND E. OJA, Optimal adaptive compression for high-dimensional data, *in* "Proceedings, 2nd Scand. Conf. on Image Analysis, June 15–17, 1981," Helsinki, Finland, pp. 152–157, 1981.
6. J. KARHUNEN AND E. OJA, New methods for stochastic approximation of truncated Karhunen–Loève expansions, *in* "Proceedings, 6th Int. Conf. on Pattern Recognition, Oct. 19–22, 1982," Munich, FRG, pp. 550–553, 1982.
7. T. KOHONEN, H. RIITTINEN, M. JALANKO, E. REUHKALA, AND S. HALTSONEN, A thousand-word recognition system based on the Learning Subspace Method and redundant hash addressing, *in* "Proceedings, 5th Int. Conf. on Pattern Recognition, Dec. 1–4, 1980," Miami Beach, Fl., pp. 158–165, 1980.
8. T. P. KRASULINA, Method of stochastic approximation in the determination of the largest eigenvalue of the mathematical expectation of random matrices, *Automat. Remote Control* **2** (1970), 215–221.
9. H. J. KUSHNER AND D. S. CLARK, "Stochastic Approximation Methods for Constrained and Unconstrained Systems," Springer-Verlag, New York, 1978.
10. M. G. LARIMORE AND R. J. CALVERT, Convergence studies of Thompson's unbiased adaptive spectral estimator, *in* "Proceedings, 14th Asilomar Conf. on Circuits, Systems, and Computers, Nov. 1980," Pacific Grove, Cal., pp. 258–262, 1980.
11. M. LOÈVE, "Probability Theory," Springer-Verlag, New York, 1977.
12. E. OJA AND J. KARHUNEN, Recursive construction of Karhunen–Loève expansions for pattern recognition purposes, *in* "Proceedings, 5th Int. Conf. on Pattern Recognition, Dec. 1–4, 1980," Miami Beach, Fl., pp. 1215–1218, 1980.
13. E. OJA AND J. KARHUNEN, An analysis of convergence for a learning version of the subspace method, *J. Math. Anal. Appl.* **91** (1983), 102–111.
14. N. L. OWSLEY, Adaptive data orthogonalization, *in* "Proceedings, 1978 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, April 10–12, 1978," Tulsa, Okl., pp. 109–112, 1978.
15. V. F. PISARENKO, The retrieval of harmonics from a covariance function, *Geophys. J. Roy. Astron. Soc.* **33** (1973), 347–366.

16. V. U. REDDY, B. EGARDT, AND T. KAILATH, Least squares type algorithm for adaptive implementation of Pisarenko's harmonic retrieval method, *IEEE Trans. Acoust. Speech Signal Proces.* **ASSP-30** (1982), 399–405.
17. H. RUTISHAUSER, Computational aspects of F. L. Bauer's simultaneous iteration method, *Numer. Math.* **13** (1969), 4–13.
18. M. SHIMURA AND T. IMAI, Nonsupervised classification using the Principal Component, *Pattern Recognition* **15** (1973), 353–363.
19. M. TASTO AND P. WINTZ, Image coding by adaptive block quantization, *IEEE Trans. Comm. Tech.* **COM-19** (1971), 957–971.
20. P. A. THOMPSON, An adaptive spectral analysis technique for unbiased frequency estimation in the presence of white noise, *in* "Proceedings, 13th Asilomar Conf. on Circuits, Systems, and Computers, Nov. 1979," Pacific Grove, Cal., pp. 529–533, 1979.
21. M. T. WASAN, "Stochastic Approximation," Cambridge Univ. Press, Cambridge, 1969.
22. H. WOLD, Estimation of principal components and related models by iterative least squares, *in* "Multivariate Analysis" (P. R. Krishnaiah, Ed.), pp. 391–420, Academic Press, New York, 1966.

Printed by the St. Catherine Press Ltd., Tempelhof 41, Bruges, Belgium