Title

Global convergence of Oja's subspace algorithm for principal component extraction

Journal

IEEE Transactions on Neural Networks, 9(1)

ISSN

1045-9227

Authors

Hua, Yingbo
Chen, Tianping
Yan, Wei-Yong

Publication Date

1998

Peer reviewed

Global Convergence of Oja's Subspace Algorithm for Principal Component Extraction

Tianping Chen, *Member, IEEE,* Yingbo Hua, *Senior Member, IEEE,* and Wei-Yong Yan

***Abstract*—Oja's principal subspace algorithm is a well-known and powerful technique for learning and tracking principal information in time series. A thorough investigation of the convergence property of Oja's algorithm is undertaken in this paper. The asymptotical convergence rates of the algorithm is discovered. The dependence of the algorithm on its initial weight matrix and the singularity of the data covariance matrix is comprehensively addressed.**

***Index Terms*—Convergence rate, global convergence, principal components extraction.**

I. Introduction

In MANY information processing systems, it is necessary to extract the main features inherent in complex high-dimensional input data streams. Such "dimensionality reduction" helps eliminate information redundancy and allows for further information transmission through limited channels. One of the most general-purpose feature extraction techniques is principal component analysis (PCA). In the unconstrained "dimensionality reduction" problem, PCA gives the estimate that best retains the information content in the mean-square sense. PCA is closely related to such concepts as Karhunen–Loeve transformation, least squares fitting, factor analysis, singular value decomposition, and matched filtering. Recently, there has been much interest in developing and studying PCA algorithms; see, e.g., [1]–[9].

One of the most important principal component extraction techniques is Oja's subspace algorithm, which is a parallel algorithm for extracting the principal subspace of a vector random process and can be implemented by linear neural networks. Moreover, under some very mild assumptions, many other algorithms can be reduced to Oja's subspace algorithm. In [5], Oja summarized various algorithms and discussed their asymptotic stability properties.

Manuscript received February 24, 1996; revised April 25, 1996, December 30, 1996, and September 4, 1997. This work was supported by the Australian Cooperative Research Center for Sensor Signal and Information Processing, the Australian Research Council, and the University of Melbourne's Collaborative Research Program. T. Chen was supported in part by the National Science Foundation of China.

T. Chen is with Department of Mathematics, Fudan University, Shanghai 200433, P.R. China.

Y. Hua is with Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052, Australia.

W.-Y. Yan is with School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.

Publisher Item Identifier S 1045-9227(98)00386-5.

The algorithm obtained by Oja in [4] can be formulated by the following iteration equation:

$$M(t)-M(t-1)=\gamma\{x(t)-M(t-1)y(t)\}y^T(t)$$
$$y(t)=M^T(t)x(t) \tag{1}$$

where $x(t)$ is the vector random process, $M(t)$ is an $n\times p$ weight matrix, and γ is a fixed number (step size). The corresponding ordinary differential equation (ODE) is

$$\frac{dM(t)}{dt}=CM(t)-M(t)M^T(t)CM(t) \tag{2}$$

where C is the covariance matrix $E(xx^T)$ with the singular value decomposition $C=U\Lambda U^T$.

The single-neuron case $(p=1)$ was considered first by Amari and Oja in [1] and [2]. The connection between the discrete algorithm and the ODE was first established by Oja and Karhunen in [3]. For the multineuron case $(p>1)$, the orthogonality of the subspace algorithm was revealed in [4]. Since then, there has been much work devoted to revealing the properties of the ODE through various approaches such as gradient descent, optimization techniques, and asymptotic analysis, etc. An explicit expression of the solution to (2) in the time domain was given in [10], which, as will be shown, is a powerful tool for further analyzing the global convergence of the subspace algorithm.

Most of the existing results do not provide the global asymptotic analysis. As Oja pointed out in his paper [5], *to analyze the global behavior for general initial conditions seems a challenging problem.* Indeed, it is of fundamental importance to develop a rigorous mathematical analysis of the global convergence of the subspace algorithm.

The purpose of this paper is to address the issue of global convergence thoroughly from a mathematical point of view. More specifically, we shall aim to gain an in-depth understanding of the following concerns:

1) the global convergence of the solution to the differential equation (2) when an arbitrary initial weight matrix is possibly rank-deficient and the covariance matrix of the inputs is possibly singular or has multiple eigenvalues;
2) the dependence of equilibria on the initial weight matrix;
3) the sharp rate of global convergence.

This paper is organized as follows. In Section II, we develop an asymptotic representation of solutions to (2) as well as some necessary notations. In Section III, we present and prove the main convergence theorem (Theorem 3). In Section IV, we give several simulations to verify the validity of the

main theorem. In Section V, we end the paper with some conclusions.

II. Representation of Solutions

Note that the covariance matrix C can be decomposed as $C = U\Lambda U^T$, where U is orthogonal and

$$\Lambda = \begin{bmatrix} \lambda_1 I_{l_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_k I_{l_k} \end{bmatrix} \quad \text{with}$$
$$\lambda_1 > \lambda_2 > \cdots, > \lambda_k \quad \text{and} \quad l_1 + \cdots + l_k = n.$$

Here, I_m denotes the $m \times m$ identity matrix.

With $W(t) = U^T M(t)$, Oja's ODE is reduced to

$$\frac{dW(t)}{dt} = (I - W(t)W^T(t))\Lambda W(t). \tag{3}$$

It is important to note that there is no loss of generality in studying the global convergence of Oja's ODE (2) through the differential equation (3). In fact, we now need to show the convergence of $W(t)$ to a matrix with proper nonzero rows.

To facilitate our subsequent discussions, we introduce the following notations, which will be used throughout the paper.

$W = W(0)$ is an initial $n \times p, (p < n)$ weight matrix with w_s row vectors.

$W_j = (w_s)_{s=\Sigma_{e=1}^{j-1} l_e+1,\cdots,\Sigma_{e=1}^{j} l_e}$ (W_j is the jth block of W, with l_j rows.)

$S_j = \text{span}\{w_s\colon\ s = 1, \cdots, \Sigma_{e=1}^{j}\ l_e\}$

$r_j = \text{rank}\,(S_j)$

m is the least integer such that $r_m = \text{rank}(W)$.

$w_{s,j}^* = w_s - P_j w_s$, where P_j is the orthogonal projector to the subspace S_j with $P_0 = 0$.

$W_j^* = (w_{s,j-1}^*)_{s=\Sigma_{e=1}^{j-1} l_e+1,\cdots,\Sigma_{e=1}^{j} l_e}$

$||E||$ is the spectral norm of the matrix E and $||E||_F$ is the Frobenius norm.

$f(t) = O(g(t))$ means that $f(t)/g(t)$ is bounded for sufficiently large $t > 0$, where $f(t)$ and $g(t)$ are positive functions.

The following result gives a general representation of the solution to the above ODE in terms of an orthogonal matrix.

Theorem A: The solution to the ODE (3) can be expressed in the form

$$W(t) = X(t)[I_p + F(t)]^{-1/2} R(t), \quad \forall t \geq 0 \tag{4}$$

where $R(t)$ is some orthogonal matrix and

$$X(t) \triangleq \begin{bmatrix} e^{\lambda_1 t} W_1 \\ \vdots \\ e^{\lambda_k t} W_k \end{bmatrix} \tag{5}$$

$$F(t) \triangleq \sum_{j=1}^{k} [\exp(2\lambda_j t) - 1] W_j^T W_j. \tag{6}$$

For the purpose of proving the global convergence, we need an asymptotic representation of the solution $W(t)$ in a more convenient form.

Consider two $n \times n$ symmetric matrices A and $A + E$. Let

$$Q = [Q_1 \quad Q_2]$$

be an orthogonal matrix such that

$$Q^T A Q = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \quad Q^T E Q = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$$

with $A_1 \in R^{n_1 \times n_1}$ where n_1 is the number of columns of Q_1. If there holds

$$\delta \triangleq \min_{\lambda_1 \in \Lambda_1, \lambda_2 \in \Lambda_2} |\lambda_1 - \lambda_2| - ||E_{11}|| - ||E_{22}|| > 2||E_{12}||$$

where Λ_i denotes the spectrum of A_i for $i = 1, 2$, then it is known from [11] that there exists a matrix $P \in R^{(n-n_1)\times n_1}$ satisfying

$$||P|| \leq 2||E_{21}||/\delta$$

such that the columns of $Q_1^* = (Q_1 + Q_2 P)(I + P^T P)^{-1/2}$ form an orthogonal basis for a subspace invariant under $A + E$. A direct application of this fact immediately leads to the following lemma, which is instrumental in deriving an alternative asymptotic representation of $W(t)$.

Lemma 1: Suppose A is a symmetric matrix of the form

$$A = VDV^T + E$$

where D is a symmetric matrix of full rank and V is such that $V^T V = I$. If

$$||D|| > 4||E||_F \tag{7}$$

then there exist a matrix V_e of the same size as V and a matrix D_e of the same size as D such that

$$AV_e = V_e D_e \quad \text{and} \quad V_e^T V_e = I \tag{8}$$

with

$$\begin{aligned} ||V_e - V|| &\leq \alpha||E|| \quad \text{and} \\ ||D_e - D|| &\leq \beta||E|| + \gamma||E||^2 \end{aligned} \tag{9}$$

where $\alpha, \beta,$ and γ are positive numbers independent of E.

In what follows, all the equalities involving t are valid only for sufficiently large t unless otherwise stated.

Theorem 1: Let $F(t)$ be as defined in (6), and let $(W_j^*)^T W_j^*$ be decomposed as

$$(W_j^*)^T W_j^* = V_j D_j V_j^T$$

for $j = 1, \cdots, m$, where D_j is a positive definite matrix and V_j is such that $V_j^T V_j = I$. Then there are two $p \times r_m$ matrices

$$\begin{aligned} V(t) &= [V_1(t) \cdots V_m(t)] \\ D(t) &= \text{diag}\,[D_1(t) \cdots D_m(t)] \end{aligned}$$

such that $F(t)$ can be expressed as

$$\begin{aligned} F(t) &= \sum_{j=1}^{m} [\exp(2\lambda_j t) - 1] W_j^T(t) W_j(t) \\ &= \sum_{j=1}^{m} [\exp(2\lambda_j t) - 1] V_j(t) D_j(t) V_j^T(t) \end{aligned} \tag{10}$$

with

$$V^T(t)V(t) = I \tag{11}$$

$$W_j^T(t) = (W_j^*)^T + \sum_{i=1}^{m} V_i \gamma_{i,j}^j(t) \tag{12}$$

$$V_j(t) = (W_j^*)^T \alpha^j(t) + \sum_{i=1}^{m} V_i \alpha_i^j(t) \tag{13}$$

where $||\alpha^j(t)||$ is bounded and

$$\gamma_{i,j}^j(t) = \begin{cases} O(e^{-2(\lambda_j - 1 - \lambda_j)t}) & \text{if } i \le j \\ O(e^{-2(\lambda_j - \lambda_i)t}) & \text{if } i > j \end{cases} \tag{14}$$

$$\alpha_i^j(t) = \begin{cases} O(e^{-2(\lambda_{j-1} - \lambda_j)t}) & \text{if } i \le j \\ O(e^{-2(\lambda_j - \lambda_i)t}) & \text{if } i > j. \end{cases} \tag{15}$$

Proof: See the Appendix. □

III. Main Results

In this section, we shall not only prove the global convergence of the subspace algorithm but also derive a tight rate of convergence. In doing so, all the notations introduced in the previous section will be adopted.

Main Theorem: Assume that $\lambda_m > 0$. Let a singular value decomposition be

$$W_j^* = U_j D_j^{1/2} V_j^T, \quad j = 1, \cdots, m$$

where D_j is a positive definite matrix, and U_j, V_j are such that

$$U_j^T U_j = I \quad \text{and} \quad V_j^T V_j = I.$$

Then with the following defintion:

$$G^T(\infty) = [V_1 U_1^T \quad \cdots \quad V_m U_m^T \quad 0_{p,n-\Sigma_{e=1}^m l_e}]$$

there is an orthogonal constant matrix R such that

$$||W(t) - G(\infty)R|| = O(e^{-\mu t}) \tag{16}$$

where μ is a rate of convergence given by

$$\mu \triangleq \min_{i=2,\cdots,m+1} \max\{\lambda_{j_i} - \lambda_i \lambda_{i-1} - \lambda_i\}$$

$$j_i \triangleq \min\{j\colon j \le i \text{ such that } W_i \in S_j\}. \tag{17}$$

Proof: Let

$$U_i(t) \triangleq e^{\lambda_i t} W_i [I_p + F(t)]^{-1/2}.$$

Because each $[\exp(2\lambda_i t) - 1] W_i^T W_i$ is nonnegative, it is easy to see that

$$\begin{aligned} ||U_i(t)|| &\le ||e^{\lambda_i t} W_i||\, ||\{I_p + [\exp(2\lambda_j t) - 1] W_j^T W_j\}^{-1/2}|| \\ &\le C \end{aligned}$$

for $i = 1, \cdots, k$, where C is a constant. It means that all $U_i(t)$ are bounded.

If $\Sigma_{e=1}^{j-1} l_e + 1 \le s \le \sum_{e=1}^{j} l_e$, then

$$w_s = w_{s,i}^* + \sum_{e=1}^{i-1} \alpha_{e,i} W_e$$

where α_e is a $1 \times (r_{i-1} - r_{i-2})$ vector. By the boundedness of $U_i(t)$ given before, it is easy to see that

$$\begin{aligned} e^{\lambda_j t} W_e [I_p + F(t)]^{-1/2} &= e^{(\lambda_j - \lambda_i)t} e^{\lambda_i t} W_e [I_p + F(t)]^{-1/2} \\ &= O(e^{(\lambda_j - \lambda_i)t}). \end{aligned}$$

Therefore, we have

$$\begin{aligned} W(t) &= Y(t)[I_p + F(t)]^{-1/2} R(t) + Z(t), \quad \forall t \ge 0 \\ &= W^*(t) + Z(t) \end{aligned} \tag{18}$$

where

$$Y(t) \triangleq \begin{bmatrix} e^{\lambda_1 t} W_1^* \\ \vdots \\ e^{\lambda_m t} W_m^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$W^*(t) = Y(t)[I_p + F(t)]^{-1/2}$$

$$Z(t) \triangleq \begin{bmatrix} Z_1(t) \\ \vdots \\ Z_k(t) \end{bmatrix}$$

and $||Z_i(t)|| = O(e^{-(\lambda_{j_i} - \lambda_i)t})$.

Without loss of generality, we shall replace $W(t)$ by $W^*(t)$. Moreover, in view of the facts that if $i > m$, then $U_i(t) = O(e^{-(\lambda_{j_i} - \lambda_i)t})$ and that if $W_i^* = 0$ for some $i < m$, then $W_i(t) = O(e^{-(\lambda_{j_i} - \lambda_i)t})$, it suffices to consider the case where $i \le m$ and $j_i = i$.

By Theorems A and 1, we have

$$\begin{aligned} W^*(t) &= Y(t)[I_p + F(t)]^{-1/2} R(t) \\ &= Y(t) \Bigg\{ I_p + \sum_{j=1}^{m} [\exp(2\lambda_j t) - 1] V_j(t) D_j(t) \\ &\qquad \cdot V_j^T(t) \Bigg\}^{-1/2} R(t) \\ &= Y(t) \Bigg\{ \sum_{j=1}^{m} [\exp(2\lambda_j t) - 1] V_j(t) D_j^*(t) V_j^T(t) \\ &\qquad + V_{m+1} V_{m+1}^T \Bigg\}^{-1/2} R(t) \\ &= Y(t) \Bigg\{ \sum_{j=1}^{m} [\exp(2\lambda_j t) - 1]^{-1/2} V_j(t) (D_j^*(t))^{-1/2} \\ &\qquad \cdot V_j^T(t) + V_{m+1}(t) V_{m+1}^T(t) \Bigg\} R(t) \end{aligned} \tag{19}$$

where $D_j^*(t) = D_j(t) + (1/[\exp(2\lambda_j t) - 1]) I_{r_j - r_{j-1}}$, and V_{m+1} is composed by $p - r_m$ normal orthogonal vectors such that $V_i^T(t) V_{m+1}(t) = 0$ for all $i = 1, \cdots, m$.

Let

$$U_i^*(t) \triangleq e^{\lambda_i t} W_i^* [I_p + F(t)]^{-1/2}$$

then

$$\begin{aligned}U_i^*(t) =& \exp(\lambda_i t)W_i^*\left\{\sum_{j=1}^{m}[\exp(2\lambda_j t)-1]^{-1/2}V_j(t)\right.\\ &\left.\cdot(D_j^*(t))^{-1/2}V_j^T(t)+V_{m+1}(t)V_{m+1}^T(t)\right\}\\ =& \exp(\lambda_i t)W_i^*\sum_{j=1}^{i-1}[\exp(2\lambda_j t)-1]^{-1/2}V_j(t)\\ &\cdot(D_j^*(t))^{-1/2}V_j^T(t)]+\exp(\lambda_i t)W_i^*\\ &\cdot\sum_{j=i+1}^{m}[\exp(2\lambda_j t)-1]^{-1/2}V_j(t)\\ &\cdot(D_j^*(t))^{-1/2}V_j^T(t)]\\ &+\exp(\lambda_i t)W_i^*[\exp(2\lambda_i t)-1]^{-1/2}V_i(t)\\ &\cdot(D_i^*(t))^{-1/2}V_i^T(t)+\exp(\lambda_i t)W_i^*\\ &\cdot V_{m+1}(t)V_{m+1}^T(t)\\ =& I_{i,1}(t)+I_{i,2}(t)+I_{i,3}(t)+I_{i,4}(t)\end{aligned} \tag{20}$$

By the similar argument used before for the case $i>m$, it is easy to see that

$$I_{i,1}(t)=O(e^{-\mu t}). \tag{21}$$

Because $W_i^*(W_j^*)^T=0, W_i^*V_j=0$ if $i\neq j$. Replacing $V_j(t)$ by $(W_j^*)^T\alpha^j(t)+\Sigma_{e=1}^{m}V_e\alpha_e^j(t)$, and noticing $||\alpha_i^j(t)||=O(e^{-2(\lambda_i-\lambda_j)t})$, we have

$$\begin{aligned}I_{i,2}(t) &= \exp(\lambda_i t)W_i^*\sum_{j=i+1}^{m}[\exp(2\lambda_j t)-1]^{-1/2}V_j(t)\\ &\quad\cdot(D_j^*(t))^{-1/2}V_j^T(t)]\\ &= \exp(\lambda_i t)W_i^*\sum_{j=i+1}^{m}[\exp(2\lambda_j t)-1]^{-1/2}V_j(t)\\ &\quad\cdot(D_j^*(t))^{-1/2}V_j^T(t)]\\ &= \exp(\lambda_i t)W_i^*\sum_{j=i+1}^{m}[\exp(2\lambda_j t)-1]^{-1/2}V_i\alpha_i^j(t)\\ &\quad\cdot(D_j^*(t))^{-1/2}V_j^T(t)]\\ &= O\left(\sum_{j=i+1}^{m}e^{\lambda_i t}e^{-\lambda_j t}e^{-2(\lambda_i-\lambda_j)t}\right)\\ &= O\left(\sum_{j=i+1}^{m}e^{-(\lambda_i-\lambda_j)t}\right)=O(e^{-\mu t})\end{aligned} \tag{22}$$

and

$$I_{i,4}(t)=0. \tag{23}$$

By the definition of $D_i^*(t)$, it is easy to see that

$$(D_i^*(t))^{-1/2}=(D_i(t))^{-1/2}+O(e^{-2\lambda_i t}).$$

Therefore

$$\begin{aligned}&V_i(t)(D_i^*(t))^{-1/2}V_i^T(t)\\ &\quad=V_i(t)(D_i(t))^{-1/2}V_i^T(t)+O(e^{-2\lambda_i t})\\ &\quad=W_i^T(t)W_i(t)+O(e^{-2\lambda_i t})\end{aligned} \tag{24}$$

then

$$\begin{aligned}I_{i,3}(t) &= \exp(\lambda_i t)W_i^*[\exp(2\lambda_i t)-1]^{-1/2}V_i(t)\\ &\quad\cdot(D_i^*(t))^{-1/2}V_i^T(t)\\ &= \exp(\lambda_i t)W_i^*[\exp(2\lambda_i t)-1]^{-1/2}V_i(t)\\ &\quad\cdot(D_i(t))^{-1/2}V_i^T(t)\\ &= \exp(\lambda_i t)W_i^*[\exp(2\lambda_i t)-1]^{-1/2}\\ &\quad\cdot(W_i^T(t)W_i(t))^{-1/2}+O(e^{-2\lambda_i t})\\ &= W_i^*((W_i^*(t))^TW_i^*(t))^{-1/2}+O(e^{-2\lambda_i t})\\ &= W_i^*((W_i^*)^TW_i^*)^{-1/2}+O(e^{-2\mu t})\\ &= U_iD_i^{1/2}V_i^TV_iD_i^{-1/2}V_i^T+O(e^{-2\mu t})\\ &= U_iV_i^T+O(e^{-2\mu t}).\end{aligned} \tag{25}$$

Therefore, we have

$$\begin{aligned}U_I^*(t)=&\left\{\sum_{j=1}^{m}[\exp(2\lambda_j t)-1]^{-1/2}V_j(t)(D_j^*(t))^{-1/2}V_j^T(t)\right.\\ &\left.+V_{m+1}(t)V_{m+1}^T(t)\right\}-U_iV_i^T=O(e^{-\mu t}).\end{aligned} \tag{26}$$

In summary, we have proved that

$$\begin{aligned}&||W(t)-G(\infty)R(t)||_F\\ &\quad=\left\|X(t)\left\{I_p+\sum_{j=1}^{k}[\exp(2\lambda_j t)-1]W_j^TW_j\right\}^{-1/2}\right.\\ &\qquad\left.\cdot\ R(t)-G(\infty)R(t)\right\|_F\\ &\quad=O(e^{-\mu t}).\end{aligned} \tag{27}$$

From the definition of $G(\infty)$, it can be verified directly that

$$[I_n-G(\infty)G^T(\infty)]\Lambda G(\infty)=\begin{pmatrix}0\\ I_{n-\Sigma_{e=1}^{m}l_e}\end{pmatrix}\Lambda G(\infty)=0. \tag{28}$$

Substituting (27) and (28) into (3) results in

$$\begin{aligned}\left\|\frac{dW(t)}{dt}\right\|_F &= ||(I-W(t)W^T(t))\Lambda W(t)-0||_F\\ &= O(e^{-\mu t})\end{aligned} \tag{29}$$

where "0's" in the previous equations are $n\times p$ zero matrices. Since

$$||W(t_2)-W(t_1)||_F=\left\|\int_{t_1}^{t_2}\dot{W}(t)\,dt\right\|_F=O(e^{-\mu(t_2-t_1)})$$

the limit of $W(t)$ exists when t tends to ∞ and thus there is a constant orthogonal matrix R such that

$$||W(t) - G(\infty)R||_F = O(e^{-\mu t}). \tag{30}$$

The theorem is proved completely.

Remark 1: The assumption of $\lambda_m > 0$ is essential for extracting principal subspace by Oja's algorithm. If $\lambda_m = 0$, then $G^T(\infty) = ([U_1V_1^T]^T, \cdots, [U_{m-1}V_{m-1}^T]^T, [W_m^{**}]^T)^T$, where $W_m^{**} = (w_s - P_{m-1}w_s)_{s=\Sigma_{e=1}^{m-1} l_e+1,\cdots,\Sigma_{e=1}^{k} l_e}$. In this case, every column vector of the limiting solution contains all components, i.e., we cannot extract principal component subspace. In fact, by noticing that $\lambda_j = 0$ for all $j \geq m$, it is easy to see that the subspace of the following:

$$W_m^{**}\left\{I_p + \sum_{j=1}^{k}[\exp(2\lambda_j t) - 1]W_j^T W_j\right\}^{-1/2} = W_m^{*}\left\{I_p + \sum_{j=1}^{m-1}[\exp(2\lambda_j t) - 1]W_j^T W_j\right\}^{-1/2} \tag{31}$$

keeps invariant and

$$Y_m^{**}\left\{I_p + \sum_{j=1}^{k}[\exp(2\lambda_j t) - 1]W_j^T W_j\right\}^{-1/2} = Y_m^{**}\left\{I_p + \sum_{j=1}^{m-1}[\exp(2\lambda_j t) - 1]W_j^T W_j\right\}^{-1/2} = O(e^{-\mu t}) \tag{32}$$

where

$$Y_m^{**} = (P_{m-1}w_s)_{s=\Sigma_{e=1}^{m-1} l_e+1,\cdots,\Sigma_{e=1}^{k} l_e}.$$

Remark 2: Theorem 3 guarantees that Oja's algorithm yields the principal subspace without the requirement that the correlation matrix C and initial weight matrix W be nonsingular.

Remark 3: Note that if C_i is an eigenspace spanned by eigenvectors corresponding to the eigenvalue λ_i, then up to an orthogonal transform the extracted component in C_i will be $U_iV_i^T$ with the number of columns being $r_{i+1} - r_i$, which is completely determined by the initial value matrix W. In addition, if $r_{i+1} - r_i$ happens to be zero, then no component in C_i can be extracted, which will be verified by simulations 2 and 3 in the following section.

Remark 4: Due to the sensitivity of the rank deficiency to small perturbation, the equilibrium approached by the solution associated with a rank-deficient initial weight matrix must be unstable. This will be verified by simulation 2 in the sequel.

Remark 5: The convergence rate is the tightest in generic case, which can been shown as following.

Suppose that $\mu = \lambda_{j_{i_0}} - \lambda_{i_0}$, for some i_0, and let

$$U_i(t) \triangleq e^{\lambda_i t}W_i[I_p + F(t)]^{-1/2}$$

as before, then

$$W_{i_0} = \sum_{e=1}^{j_{i_0}} \alpha_{e,i_0}W_e$$

and

$$U_{i_0}(t) = e^{\lambda_{i_0}t}\sum_{e=1}^{j_{i_0}} \alpha_{e,i_0}W_e[I_p + F(t)]^{-1/2} = \sum_{e=1}^{j_{i_0}} \alpha_{e,i_0}e^{(\lambda_{i_0}-\lambda_e)t}U_e(t).$$

Therefore, we can find a constant D such that for sufficiently large t

$$\frac{||U_{i_0}(t)||_F}{e^{(\lambda_{i_0}-\lambda_{j_{i_0}})t}} \geq ||\alpha_{j_{i_0},i_0}U_{j_{i_0}}(t)||_F - \left|\left|\sum_{e=1}^{j_{i_0}-1} \alpha_{e,i_0}e^{(\lambda_{j_{i_0}}-\lambda_e)t}U_e(t)\right|\right|_F > D$$

which, combining with the main theorem, leads to that there are two constants D_1, D_2 such that

$$D_1e^{-\mu t} \leq ||W(t) - W(\infty)R||_F \leq D_2e^{-\mu t} \tag{33}$$

and the convergence rate $e^{-\mu t}$ is the tightest in the generic case.

Theorem 3: Under the assumptions that the first p rows of $W(0)$ are linearly independent and there is an index m such that

$$\sum_{i=1}^{m} l_i = p \quad \text{and} \quad \lambda_{m+1} > 0$$

there holds

$$||W(t) - G(\infty)R|| = O(e^{-(\lambda_m - \lambda_{m+1})t}). \tag{34}$$

Proof: See the Appendix. □

IV. Simulations

In this section, a number of simulation results will be reported. The purpose here is to verify the theoretical results obtained in the paper and support some conclusions we draw. Additional interesting phenomena will also be observed.

All the subsequent figures will contain plots of the derivative of

$$\ln ||W(t) - \lim_{t\to\infty} W(t)||$$

versus t since this derivative serves as a meaningful measure of the convergence rate. In fact, the negative of the derivative will be called the convergence rate.

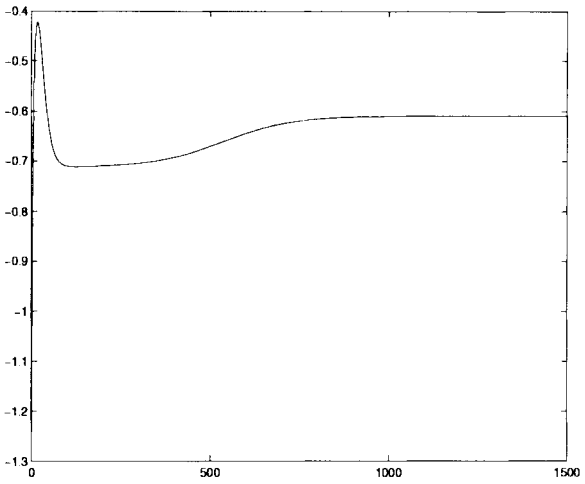

Fig. 1. Simulation 1 (Negative convergence rate versus iterations).

Simulation 1: This simulation verifies the validity of Theorem 3 even when C has multiple eigenvalues.

Let the covariance matrix be

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{35}$$

and the initial weight matrix be

$$W(0) = \begin{pmatrix} 1.1650 & 1.6961 & -1.4462 \\ 0.6268 & 0.0591 & -0.7012 \\ 0.0751 & 1.7971 & 1.2460 \\ 0.3516 & 0.2641 & -0.6390 \\ -0.6965 & 0.8717 & 0.5774 \end{pmatrix} \tag{36}$$

where the first three rows of $W(0)$ are independent. Taking $\gamma = 0.05$ and running 500 iterations, we have

$$W(500) = \begin{pmatrix} 0.2988 & 0.6705 & -0.6791 \\ 0.9420 & -0.3210 & 0.0975 \\ 0.1526 & 0.6688 & 0.7276 \\ 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 \end{pmatrix}. \tag{37}$$

Clearly, $W(t)$ is seen (or easy to verify) to converge to an orthogonal matrix, whose columns are linear combinations of the eigenvectors corresponding to the first three largest eigenvalues, i.e., they are located in the three-dimensional principal eigenspace of the covariance matrix. Fig. 1 shows that the convergence rate is close to $\lambda_2 - \lambda_3 = 0.6$, which agrees with Theorem 4 (with $m = 2$).

Simulations 2 and 3 will show what happens when the initial matrix is of rank deficiency or $\lambda_m = 0$.

Simulation 2: Let

$$\Lambda = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{38}$$

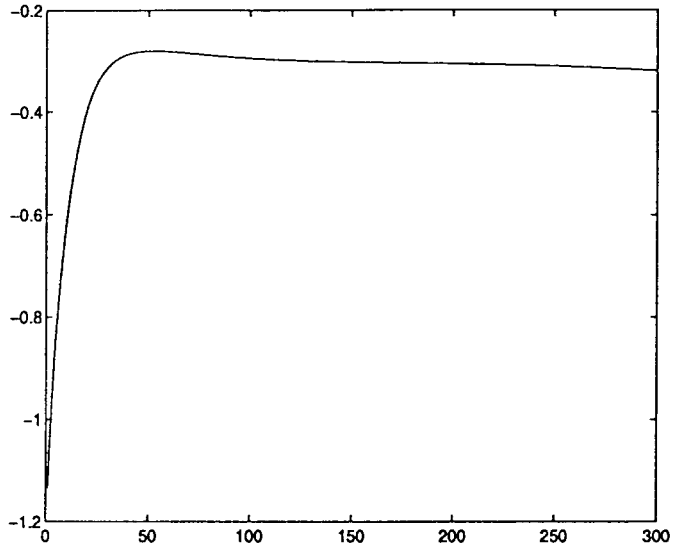

Fig. 2. Simulation 2: (negative) convergence rate versus the first 300 iterations.

and

$$W(0) = \begin{pmatrix} 0.1551 & 0.1302 & -1.0038 \\ -0.9646 & 0.0350 & -0.4974 \\ -0.9646 & 0.0350 & -0.4974 \\ 0.1551 & 0.1302 & -1.0038 \\ -0.9646 & 0.0350 & -0.4974 \end{pmatrix} \tag{39}$$

respectively, where $\operatorname{rank}(W(0)) = 2$. In this case, the first two rows of $W(0)$ span the row space of $W(0)$.

Taking $\gamma = 0.05$ and running 1500 iterations, we have

$$W(1500) = \begin{pmatrix} 0.2793 & 0.1272 & -0.9517 \\ -0.9598 & 0.0080 & -0.2806 \\ -0.0005 & 0.0000 & -0.0001 \\ -0.0000 & 0.0000 & 0.0000 \\ -0.0000 & 0.0000 & 0.0000 \end{pmatrix} \tag{40}$$

which shows that $W(500)$ extracts the principal subspace spanned by the two eigenvectors corresponding to the first two largest eigenvalues. Fig. 2 shows that the convergence rate is about $\mu = \lambda_2 - \lambda_3 = 0.3$, the smallest among $\lambda_i - \lambda_{i+1}$, which confirms Theorem 3.

However, if we iterate 1000 more times, the final result becomes

$$W(1500) = \begin{pmatrix} 0.2793 & 0.1272 & -0.9517 \\ -0.9598 & 0.0080 & -0.2806 \\ -0.0281 & 0.9918 & 0.1243 \\ -0.0000 & 0.0000 & 0.0000 \\ -0.0000 & 0.0000 & 0.0000 \end{pmatrix} \tag{41}$$

which shows that $W(1500)$ extracts the principal subspace spanned by the three eigenvectors corresponding to the first three largest eigenvalues.

This phenomenon does not violate our theory. The reason for this is that when we iterate, the numerical rounding errors make $W(t)$ become of full rank. Therefore, the dimension of the principal subspace becomes three. All these results coincide with the conclusion of our main theorem. It also means that the theoretical equilibrium due to rank deficient initial matrix $W(0)$ is unstable. Fig. 3 shows the last 600 iterations leading to $W(1500)$. It can be seen that the convergence rate

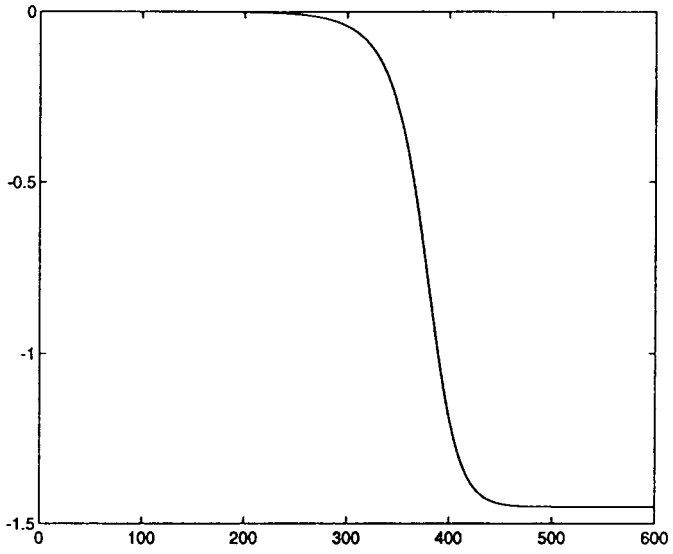

Fig. 3. Simulation 2: (negative) convergence rate versus the last 600 iterations of total 1500 iterations.

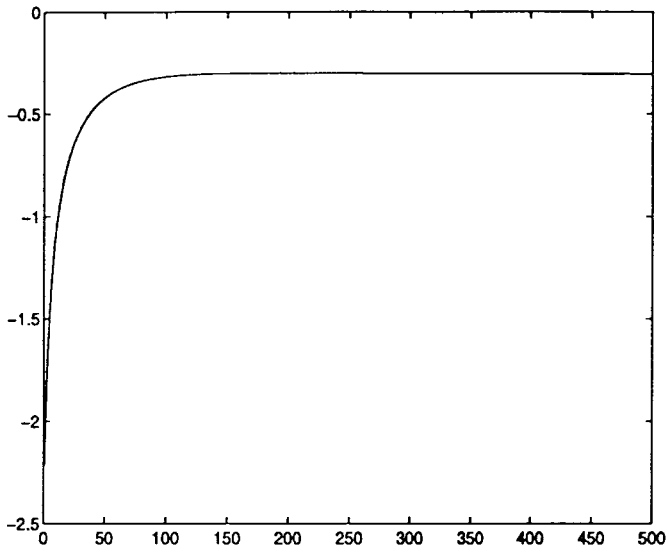

Fig. 4. Simulation 3: (negative) convergence rate versus iterations.

is nearly zero initially in the figure (equilibrium) and then becomes$2\lambda_m$ after the random numerical rounding error makes $W(t)$ becomes effectively full rank.

Simulation 3: This simulation shows that in some cases, the eigenspace corresponding to a larger eigenvalue may be eliminated while the eigenspace corresponding to a smaller eigenvalue may be extracted, as indicated in Remark 3.

Let

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{42}$$

and

$$W(0) = \begin{pmatrix} 1.1650 & 1.6961 & -1.4462 \\ 1.1650 & 1.6961 & -1.4462 \\ 0.0751 & 1.7971 & -1.2460 \\ 0.086\,42 & 1.5831 & -1.4763 \\ -0.6965 & 0.8717 & 0.5774 \end{pmatrix} \tag{43}$$

respectively, where $m = 3$ and $\lambda_m = 0$.

Taking $\gamma = 0.03$ and running 1000 iterations, we have

$$W(1000) = \begin{pmatrix} 0.4684 & 0.6142 & -0.6351 \\ 0.0001 & -0.0005 & -0.0006 \\ -0.0406 & 0.7330 & 0.6790 \\ 0.1389 & -0.0460 & 0.0579 \\ -0.5797 & 0.1919 & -0.2419 \end{pmatrix}. \tag{44}$$

In this simulation, the second row of $W(0)$ is linearly dependent on the first row, and $r_1 = 1, r_2 = 2, r_3 = 3, w_{2,1}^* = 0$. It is clear from the above matrix that the first row of the second block (corresponding to λ_2) in the limiting solution $W(1000)$ is approximately zero and the rows of the third block (corresponding to $\lambda_3 = 0$) still remain. This justifies the conclusion of our main theorem and our remark 3. Moreover, the convergence rate μ is determined by the minimum of $1 - 0.7$ and $0.7 - 0$, which can be seen from Fig. 4.

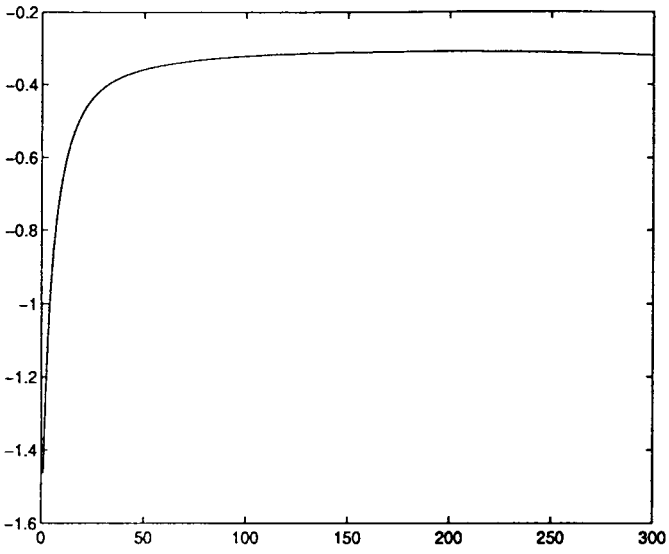

Fig. 5. Simulation 4: (negative) convergence rate versus interations.

It is worth pointing out that in this simulation, $\lambda_m = 0$, where $m = 3$,which causes that $W^T(1000)W(1000)$ is not an orthogonal matrix.

Simulation 4: Let the covariance matrix and the initial weight matrix be

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{45}$$

and

$$W(0) = \begin{pmatrix} 1.1650 & 1.6961 & -1.4462 \\ 1.1650 & 1.6961 & -1.4462 \\ 0.0751 & 1.7971 & 1.2460 \\ 0.0751 & 1.7971 & 1.2460 \\ 0.0751 & 1.7971 & 1.2460 \end{pmatrix} \tag{46}$$

respectively.

Taking $\gamma = 0.05$ and running 500 iterations, we have

$$W(500) = \begin{pmatrix} 0.4687 & 0.6086 & -0.6403 \\ 0.0003 & -0.0012 & -0.0017 \\ -0.0370 & 0.7377 & 0.6741 \\ -0.0000 & 0.0000 & 0.0000 \\ -0.0000 & 0.0000 & 0.0000 \end{pmatrix}. \quad (47)$$

In this simulation, $r_1 = 1, r_2 = 2, r_3 = r_4 = 2, w^*_{2,1} = 0, w^*_{4,2} = 0, w^*_{5,3} = 0$, therefore, the second, fourth, and fifth rows of $W(500)$ are approximately zero, which coincides with the conclusion of our main theorem.

It is interesting to observe that due to the rounding errors, starting from any $n \times p$ initial matrix $W(0)$, the solution $W(t)$ always converges to an $n \times p$ full rank matrix with the upper submatrix being an orthogonal matrix and the lower submatrix being zero. Furthermore, after $W(t)$ becomes full rank, the convergence rate is given by Theorem 4.

V. Conclusions

In this paper, we have analyzed in detail the dynamic behavior of Oja's subspace algorithm. The global convergence of the algorithm has been established and its convergence rate has been found. The dependence of the convergence on the initial value matrix has been explored completely. In particular, we have discussed thoroughly the case when the autocorrelation matrix has multiple eigenvalues or is singular. All these results are summarized in the Main Theorem. We have also demonstrated that in practice (with random rounding error), starting from any initial p-column weight matrix, the algorithm will eventually extract a p-dimensional principal subspace with a global convergence rate given by Theorem 3. In the practical sense, therefore, Main Theorem provides possible transient convergence rate which could be much slower than the final convergence rate given by Theorem 3.

Appendix

Proof of Theorem 1: Let

$$F_{11}(t) \triangleq V_1 D_1 V_1^T + \frac{\exp(2\lambda_2 t) - 1}{\exp(2\lambda_1 t) - 1} W_2^T W_2$$

by Lemma 1, we have

$$F_{11}(t) V_{11}(t) = V_{11}(t) D_{11}(t).$$

Because all the row vectors of W_2 belong to S_2, all the column vectors in $V_{11}(t)$ belong to S_2, too. Thus we have

$$V_{11}(t) = V_1 + V_1 \alpha_{11}(t) + V_2 \alpha_{12}(t)$$
$$D_{11}(t) = \beta_{11}(t) + D_1$$

where for $i = 1, 2 \alpha_{1i}(t)$ are $(r_i - r_{i-1}) \times r_i$ matrices and $\beta_{11}(t)$ is an $r_1 \times r_1$ matrix with

$$||\alpha_{1i}(t)|| = O(e^{-(2\lambda_1 - 2\lambda_2)t}) \qquad i = 1, 2$$
$$||\beta_{11}(t)|| = O(e^{-(2\lambda_1 - 2\lambda_2)t}).$$

Letting

$$F_{1,i+1}(t) \triangleq F_{1,i}(t) + \frac{\exp(2\lambda_{i+1} t) - 1}{\exp(2\lambda_1 t) - 1} W_{i+1}^T W_{i+1}$$

and repeating this procedure, by induction we have

$$V_1(t) = V_1 + V_1 \alpha_1^1(t) + \cdots + V_m \alpha_m^1(t)$$
$$D_1(t) = D_1 + \beta_1(t)$$
$$F(t) V_1(t) = V_1(t) D_1(t)$$

and

$$||\alpha_1^1(t)|| = O(e^{-(2\lambda_1 - 2\lambda_2)t})$$
$$||\alpha_i^1(t)|| = O(e^{-(2\lambda_1 - 2\lambda_i)t})$$
$$||\beta_1(t)|| = O(e^{-(2\lambda_1 - 2\lambda_2)t})$$

for $i = 1, \cdots, m$, where $\alpha_i^1(t)$ are $(r_i - r_{i-1}) \times r_1$ matrices and $\beta_1(t)$ is a $r_1 \times r_1$ matrix.

With

$$W_{j,1}^T(t) = W_j^T - V_1(t) V_1^T(t) W_j^T$$

it is easy to check that there are $\gamma_{i,j}^1(t)$ such that

$$W_{j,1}^T(t) = W_j^T - V_1 V_1^T W_j^T + \sum_{i=1}^m V_i \gamma_{i,j}^1(t)$$
$$W_{1,1}^T(t) = \sum_{i=1}^m V_i \gamma_{i,1}^1(t)$$

where

$$||\gamma_{i,j}^1(t)|| = O(e^{-2(\lambda_1 - \lambda_i)t}) \qquad i = 1, \cdots, m, \quad j = 1, \cdots, k$$

and

$$F(t) = [e^{2\lambda_1 t} - 1] W_1^T(t) W_1(t) + \sum_{j=1}^k [e^{2\lambda_j t} - 1] W_{j,1}^T(t) W_{j,1}(t)$$

where

$$W_1(t) = W_1 - W_{1,1}(t).$$

In a similar way, by letting

$$F_{2,0}(t) = W_{2,2}^T(t) W_{2,2}(t) + [\exp(2\lambda_2 t) - 1] W_{2,1}^T(t) W_{2,1}(t)$$

and

$$F_{i+1,2}(t) \triangleq F_{i,2}(t) + \frac{\exp(2\lambda_{i+1} t) - 1}{\exp(2\lambda_2 t) - 1} W_{i+1,1}^T(t) W_{i+1,1}(t)$$

we have

$$W_{j,2}^T(t) = W_j^T - [V_1, V_2][V_1, V_2]^T W_j^T + \sum_{i=1}^m V_i \gamma_{i,j}^2(t)$$
$$W_{2,2}^T(t) = \sum_{i=1}^m V_i \gamma_{i,2}^2(t)$$

where

$$||\gamma_{i,2}^2(t)|| = \begin{cases} O(e^{-2(\lambda_1 - \lambda_2)t}) & \text{if } i = 1, 2 \\ O(e^{-2(\lambda_2 - \lambda_i)t}) & \text{if } i > 2 \end{cases}$$

and

$$F(t) = \sum_{j=1}^2 [e^{2\lambda_j t} - 1] W_j^T(t) W_j(t) + \sum_{j=2}^k [e^{2\lambda_j t} - 1] \cdot W_{j,2}^T(t) W_{j,2}(t)$$
$$W_2(t) = W_2^* - W_{2,2}(t).$$

Continuing with the same argument successively m times, we have

$$F(t) = \sum_{j=1}^{m} [e^{2\lambda_j t} - 1] W_j^T(t) W_j(t)$$

where

$$W_j(t) = W_j^* - W_{j,j}(t)$$
$$W_{j,j}^T(t) = \sum_{i=1}^{m} V_i \gamma_{i,j}^j(t)$$

with

$$||\gamma_{i,j}^j(t)|| = \begin{cases} O(e^{-2(\lambda_{j-1}-\lambda_j)t}) & \text{if } i \le j \\ O(e^{-2(\lambda_j-\lambda_i)t}) & \text{if } i > j. \end{cases}$$

By the definition

$$W_j^T(t) W_j(t) = V_j(t) D_j(t) V_j^*(t)$$
$$W_j^{*^T} W_j^* = V_j D_j V_j^*.$$

Because each column vector in $V_j(t)$ is some linear combination of the column vectors of $W_j^T(t)$, therefore, there exist $\alpha^j(t), \alpha_i^j(t)$, such that

$$V_j(t) = (W_j^*)^T \alpha^j(t) + \sum_{i=1}^{m} V_i \alpha_i^j(t)$$

with $||\alpha^j(t)||$ being bounded and

$$\alpha_i^j(t) = \begin{cases} O(e^{-2(\lambda_{j-1}-\lambda_j)t}) & \text{if } i \le j \\ O(e^{-2(\lambda_j-\lambda_i)t}) & \text{if } i > j. \end{cases} \tag{48}$$

Theorem 2 is proved.

Proof of Theorem 3: With

$$X_1(t) = [W_1^T(t), \cdots, W_m^T(t)]^T$$
$$X_2(t) = [W_{m+1}^T(t), \cdots, W_k^T(t)]^T$$
$$\Lambda_1 = \text{diag}\,[\lambda_1, \cdots, \lambda_m], \quad \Lambda_2 = \text{diag}\,[\lambda_{m+1}, \cdots, \lambda_k]$$

Oja's ODE becomes

$$\begin{bmatrix} \dot{X}_1(t) \\ \dot{X}_2(t) \end{bmatrix} = \begin{bmatrix} \Lambda_1 X_1(t) \\ \Lambda_2 X_2(t) \end{bmatrix} - \begin{bmatrix} X_1(t) \\ X_2(t) \end{bmatrix} [X_1^T(t) \Lambda_1 X_1^T(t) + X_2^T(t) \Lambda_2 X_2^T(t)]. \tag{49}$$

From the proof of Theorem 3, it follows that:

$$||X_2(t)||_F = O(e^{-(\lambda_m - \lambda_{m+1})t}) \tag{50}$$

which implies that

$$||I_p - W^T(t) W(t)||_F^2 = ||I_p - X_1^T(t) X_1(t)||_F^2 + O(e^{-2(\lambda_m - \lambda_{m+1})t}) \tag{51}$$

due to

$$W^T(t) W(t) = X_1^T(t) X_1(t) + X_2^T(t) X_2(t).$$

Thus, it is shown that $X_1(t)$ converges to an orthogonal matrix, Consequently, there is $t_0 > 0$ such that there holds $X_1^T(t) X_1(t) > (1 - \lambda_{m+1}/\lambda_m) I_p$ for all $t > t_0$, leading to

$$X_1^T(t) \Lambda_1 X_1(t) > (\lambda_m - \lambda_{m+1}) I_p, \quad \forall t > t_0. \tag{52}$$

Setting

$$u(t) = ||I_p - X_1^T(t) X_1(t)||_F^2 \tag{53}$$

we see that

$$\begin{aligned} \dot{u}(t) &= -4\,\text{tr}\,[(I_p - X_1^T(t) X_1(t)) X_1^T(t) \Lambda_1 X_1(t) \cdot (I_p - X_1^T(t) X_1(t))] \\ &\quad + 4\,\text{tr}\,[X_2^T(t) \Lambda_2 X_2(t) X_1^T(t) X_1(t)] \\ &\le -4(\lambda_m - \lambda_{m+1}) u(t) + a e^{-2(\lambda_m - \lambda_{m+1})t} \\ &\quad \forall t > t_0 \end{aligned} \tag{54}$$

where the last inequality follows from (52) and (50). Quite obviously, this inequality is equivalent to

$$(e^{4(\lambda_m - \lambda_{m+1})t} u(t))' \le a e^{2(\lambda_m - \lambda_{m+1})t}$$

from which it is deduced that

$$u(t) = O(e^{-2(\lambda_m - \lambda_{m+1})t})$$

implying that

$$||I_p - X_1(t) X_1^T(t)||_F^2 = O(e^{-2(\lambda_m - \lambda_{m+1})t}). \tag{55}$$

Now from (49), it follows that:

$$\dot{X}_1(t) = (I_p - X_1(t) X_1^T(t)) \Lambda_1 X_1(t) - X_1(t) X_2^T(t) \Lambda_2 X_2(t). \tag{56}$$

This in combination with (50) and (55) yields

$$||\dot{X}_1(t)||_F = O(e^{-(\lambda_m - \lambda_{m+1})t}). \tag{57}$$

In view of

$$||X_1(t_2) - X_1(t_1)||_F = \left|\left| \int_{t_1}^{t_2} \dot{X}_1(t)\, dt \right|\right|_F$$

one sees by the Cauchy criterion that $X_1(t)$ converges as $t \to \infty$ at the exponential rate of $\lambda_m - \lambda_{m+1}$. Overall, it is concluded that

$$||W(t) - G(\infty) R|| = O(e^{-(\lambda_m - \lambda_{m+1})t}) \tag{58}$$

as required.

ACKNOWLEDGMENT

The authors would like to acknowledge the helpful comments by reviewers, which have helped to improve the readability of this paper.

REFERENCES

[1] S. I. Amari, "Neural theory of association and concept-formation," *Biol. Cybern.*, vol. 26, pp. 175–185, 1977
[2] E. Oja, "A simplified neuron model as a principal component analyzer," *J. Math. Biol.*, vol. 15, pp. 267–273, 1982.

[3] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *J. Math. Anal. Applicat.*, vol. 106, pp. 69–84, 1985.
[4] E. Oja, "Neural networks, principal components, and subspaces," *Int. J. Neural Syst.*, vol. 1, pp. 61–68, 1989.
[5] ———, "Principal components, minor components, and linear neural networks," *Neural Networks*, vol. 5, pp. 927–935, 1992.
[6] K. Hornik and C. M. Kuan, "Convergence analysis of local feature extraction algorithms," *Neural Networks*, vol. 5, pp. 229–240, 1992.
[7] P. Baldi and K. Hornik, "Learning in linear neural networks: A survey," *IEEE Trans. Neural Networks*, vol. 6, pp. 837–858, 1995.
[8] L. Xu, L. Krzyzak, and E. Oja, "Neural nets for dual subspace pattern recognition method," *Int. J. Neural Syst.*, vol. 2, pp. 169–184, 1991.
[9] L. Xu, "Least mean square error recognition principle for self-organizing neural nets," *Neural Networks*, vol. 6, pp. 627–648, 1993.
[10] W.-Y. Yan, U. Helmke, and J. B. Moore, "Global analysis of Oja's flow for neural networks," *IEEE Trans. Neural Networks*, vol. 5, pp. 674–683, 1994.
[11] G. H. Golub and Van Loan, *Matrix Computation*. Baltimore, MD: John Hopkins Univ. Press, 1985, p. 271.

Tianping Chen (SM'96) received the postgraduate degrees from Fudan University, Shanghai, China, in 1966.

He is a Professor at Fudan University as well as a concurrent Professor at Nanjing University of Aeronautics and Astronautics. He has held appointments at several institutes in the United States, Europe, Japan, and Australia. His research interests include harmonic analysis, approximation theory, neural networks, and signal processing. He has published more than 100 journal papers.

Dr. Chen was a recipient of National Awards for Excellence in Scientific Research by State Education Commission of China in 1985 and 1994.

Yingbo Hua (S'86–M'88–SM'92) was born in China in 1960. He received the B.S. degree from the Nanjing Institute of Technology (currently Southeast University), Nanjing, China, in February 1982, and the M.S. and Ph.D. degrees from Syracuse University, Syracuse, NY, in 1983 and 1988, respectively.

Since February 1990, he has been with the University of Melbourne, Australia, where he was Lecturer from 1990 to 1992, Senior Lecturer from 1993 to 1995, and has been Associate Professor and Reader since January 1996. He has published more than 50 journal papers and 80 conference papers in the areas of spectral estimation, array processing, radar and NMR imaging, system identification, neural networks, and wireless communications.

Dr. Hua has served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING since 1994 and a member of a number of international technical committees.

Wei-Yong Yan was born in Fuzhou, China. He received the B.S. degree in mathematics from Nankai University, Tianjin, China, in 1983, the M.S. degree in systems science from Academia Sinica, Beijing, China, in 1986, and the Ph.D. degree in systems engineering from the Australian National University, Canberra, in 1990.

Following the completion of his Ph.D. studies, he worked as a Research Fellow in the Department of Systems Engineering, the Australian National University. He was a Lecturer in Applied Mathematics at the University of Western Australia for two years from 1993 to 1994. Since 1995, he has been a Lee Kuan Yew Fellow in the School of Electrical and Electronic Engineering at the Nanyang Technological University in Singapore. His current research interests include signal processing, neural networks, and optimization.