Infinity Norm Based Neural Network Algorithm for Principal Component Analysis

Lijun Liu
Department of Mathematics
Dalian Nationalities University
Dalian, 116600, China

Hongjie Xing
College of Mathematics and Computer Science
Hebei University
Baoding, 071002, China

Dong Nan
College of Applied Science
Beijing University of Technology
Beijing, 100022, China

***Abstract*— In this paper a simple infinity norm based neural network algorithm for estimation of the principal component is developed. It seems to be especially useful in applications with changing environment, where the learning process has to be repeated in on-line manner. Theoretical analysis shows the weight vector converges to the principal eigenvector asymptotically. In comparison with the existing algorithms, numerical simulation shows that the proposed algorithm demonstrates fast convergence and robustness for a slightly noisy Gaussian samples with some points having large magnitude and angle with respect to the principal direction.**

***Index Terms*— Principal component analysis; Eigenvalue; Neural Network; Convergence.**

I. Introduction

Principal component analysis (PCA) is a widely used statistical technique in various applications such as feature extraction in pattern recognition, data compression and coding in signal processing. In stationary case, most adaptive algorithms are studied by explicitly solving the system of differential equations due to the fundamental stochastic approximation theorem [1][2].

Since Oja's pioneer founding [3] (referred as **OJA rule**) that a simple linear neuron with a constrained Hebbian learning rule can extract the principal component from stationary input data, there were increasing interests in the study of connections between PCA and neural networks [4][5][6][7][8][9][10]. Given a zero mean stationary process $\{x(t) \in R^n\}_{t=0}^{\infty}$ with symmetric correlation matrix $R = E(x(t)x^T(t)) \in R^{n\times n}$, where $E(\cdot)$ is the expectation operator on this stationary process. Then the adaptive OJA rule for a simple linear neuron with weight vector $w(t) = [w_1(t), w_2(t), \cdots, w_n(t)]^T \in R^n$ is as follows

$$w(t+1) = w(t) + \eta(t)[x(t)y(t) - y^2(t)w(t)], \qquad (1)$$

where index $t = 0, 1, \cdots$, $y(t) = w^T(t)x(t)$ is the linear output with current sample input $x(t)$, and the learning rate $\eta(t) > 0$ should be chosen to balance the stability property and the convergence speed of the proposed algorithm.

The term $x(t)y(t)$ is always referred to Hebbian rule. To avoid inevitable unboundedness of $w(t)$, a degenerated term $-y^2(t)$ is added, which plays a central role in controlling the stability of the OJA rule. Taking expectation on both sides of (1), we can obtain

$$w(t+1) = w(t) + \eta(t)[Rw(t) - w^T(t)Rw(t)w(t)] \qquad (2)$$

with index $t = 0, 1, \cdots$. Its stability property can be studied by explicitly solving the following ordinary differential equations (ODE) due to the fundamental stochastic approximation theorem [1][2].

$$\frac{dw(t)}{dt} = Rw(t) - w^T(t)Rw(t)w(t) \qquad (3)$$

It has been proved that eigenvector associated with the largest eigenvalue of symmetric positive matrix R is asymptotically stable [3].

Another normalized OJA type algorithm for extracting principal component is proposed by Xu et al. [8] (often referred as **OJAN rule**), which can be summarized as the following ODE

$$\frac{dw(t)}{dt} = Rw(t) - \frac{w^T(t)Rw(t)}{w^T(t)w(t)}w(t). \qquad (4)$$

Luo et al. [7][9] proposed another algorithm (referred as **LUO rule**) to compute principal component

$$\frac{dw(t)}{dt} = [w^T(t)w(t)]Rw(t) - [w^T(t)Rw(t)]w(t), \qquad (5)$$

which is further studied by Zhang et al. [4]. There, analytical solution of (5) is given and its asymptotic behavior to eigenvectors corresponding to largest eigenvalues is successfully obtained.

Liu et al. [5] proposed a simpler model (referred as **2-norm rule**) as

$$\frac{dw(t)}{dt} = Aw(t) - \left[w^T(t)w(t)\right]w(t). \qquad (6)$$

Apparently non-negative continuous function of $w^T(t)w(t) = \|w(t)\|_2^2$, where $\|\cdot\|_2$ denotes the usual Euclidean $2-$norm. Thus, it is easy to see that this term also acts as preventing trajectory $w(t)$ tending to infinity. So, many other norm choice of $w(t)$ is hopeful. For example norm with respect to a symmetric positive definite matrix $B \in R^{n\times n}$ is given as (referred as $B-$**norm rule**)

$$\frac{dw(t)}{dt} = Aw(t) - \left[w^T(t)Bw(t)\right]w(t), \qquad (7)$$

which was discussed in literature [1].

Recently a novel model is introduced [6] to compute the principal component of R as

$$\frac{dw(t)}{dt} = Rw(t) - \text{sgn}(w^T(t))w(t)w(t), \tag{8}$$

where $\text{sgn}(w^T(t))w(t)$ is equivalent to $1-$norm of $w(t)$ by introducing $\|w(t)\|_1 = \sum_{i=1}^n |w_i(t)|$. So it is equivalent to (referred as **1-norm rule**)

$$\frac{dw(t)}{dt} = Rw(t) - \|w(t)\|_1 w(t). \tag{9}$$

From the consideration of implementation for $w(t)$, the most appropriate and cheapest way should be the infinity norm $\|w(t)\|_\infty = \max_{1\le i\le n} |w_i(t)|$. Therefore, we propose the same single linear neuron but with learning rule based on infinity norm (referred as $\infty-$**norm rule**) for extracting principal component as follows

$$\frac{dw(t)}{dt} = Rw(t) - \|w(t)\|_\infty w(t). \tag{10}$$

Its adaptive form for a stationary process $\{x(t), t = 0, 1, 2, \cdots\}$ with correlation matrix R takes

$$w(t+1) = w(t) + \eta(t)\Big[x(t)y(t) - \|w(t)\|_\infty w(t)\Big], \tag{11}$$

where index $t = 0, 1, \cdots$.

As for the continuous ODE models mentioned above, theoretically analysis has shown that they are all computationally equivalent. However, they behave great differences when applied in an adaptive manner. The main reasons are twofold below.

1) The learning rate $\eta(t)$ should be selected carefully. Any inappropriate choice of it may bring about instability of the adaptive algorithms.
2) Possibility for some x having large magnitude and angle with respect to the direction of $w(t)$ also could lead to instability to the learning rules.

In this paper, one experiment is designed to demonstrate these two problems. The numerical result shows that the norm based adaptive algorithms are much stable than the OJA type algorithms. Moreover, among the three norm based algorithms, the proposed model (10) is obviously the simplest because the degeneration term only use one component of $w(t) \in R^n$. Thus, when built into hardware, the proposed infinity norm is more computationally efficient and simpler.

II. Convergence Analysis

As $\max_{1\le i\le n} |w_i|$ is not differentiable, the theory of unique existence is not so obvious for equation (10). We provide the proof for it in the following. Assume R is a symmetric positive matrix with eigenvalues $\lambda_1 > \lambda_2 \ge \cdots \ge \lambda_n > 0$ and their corresponding orthonormal eigenvectors are $S_1, S_2, \cdots, S_n$, respectively, then we have

Theorem 1: For any given $\xi \in R^n$, there exists unique solution $w = w(t)$ with initial condition $w(0) = \xi$ for equation (10).

Proof: Denote the right side of equation (10) by $f(w)$. For simplicity, in the following $\|\cdot\|_2$ is abbreviated to $\|\cdot\|$. Since $\max_{1\le i\le n} |w_i|$ is continuous on domain $\{(t, w)|0 \le t < \infty, w \in R^n\}$, for a locally bounded closed domain $\mathcal{N} \subseteq R^n$, there exists a constant $K_1 > 0$ such that $\|w(t)\|_\infty = \max_{1\le i\le n} |w_i| \le K_1$. For any $u = (u_1, \cdots, u_n)^T$ and $v = (v_1, \cdots, v_n)^T \in \mathcal{N}$, denote $\max_{1\le i\le n} |u_i| = |u_1|$ and $\max_{1\le i\le n} |v_i| = |v_1|$, we have $|u_1| \le K_1$ and $|u_2| \le K_1$. Moreover, due to the fact that all the norms in finite dimensional space are equivalent, there exists a constant $K_2 > 0$ such that $\|u\|_\infty \le K_2\|u\|$. Then we have

$$\begin{aligned}
\|f(u) &- f(v)\| \\
&\le \big\|R(u-v)\big\| + \big\||u_1|u - |v_1|v\big\| \\
&\le \big\|R\big\|\big\|u-v\big\| + \big\||u_1|u - |u_1|v\big\| + \big\||u_1|v - |v_1|v\big\| \\
&\le (|\lambda_{\max}| + K_1)\big\|u-v\big\| + (|u_1| - |v_1|)\big\|v\big\| \\
&\le (|\lambda_{\max}| + K_1)\big\|u-v\big\| + K_1|u_1 - v_1| \\
&\le (|\lambda_{\max}| + K_1)\big\|u-v\big\| + K_1\|u-v\|_\infty \\
&\le (|\lambda_{\max}| + K_1)\big\|u-v\big\| + K_1K_2\big\|u-v\big\| \\
&= (|\lambda_{\max}| + K_1 + K_1K_2)\big\|u-v\big\|,
\end{aligned} \tag{12}$$

which means that $f(w)$ is locally Lipschitz continuous. From the existence theory of ordinary differential equations [11], there exists a unique solution $w = w(t)$ with $w(0) = \xi$ for (10) on some interval $[0, T]$. ∎

Theorem 2: Let $w = w(t)$ on $0 \le t \le T$ be a solution for (10), then $\|x(t)\|$ is bounded, i.e.,

$$\min\{\|w(0)\|, \lambda_{\min}\} \le \|w(t)\| \le \max\{\|w(0)\|, \sqrt{n}\lambda_{\max}\} \tag{13}$$

for all $t \in [0, T]$. Therefore, $w(t)$ can be extended to the infinite time interval $[0, \infty)$.

Proof: Consider nonnegative function

$$E(t) = \frac{1}{2}\|w(t)\|^2. \tag{14}$$

Using the chain rule, we obtain the derivative of $E(t)$ with respect to t along trajectory of (10) as follows

$$\frac{dE}{dt} = w^T\frac{dw}{dt} = w^TRw - \|w\|_\infty w^Tw. \tag{15}$$

Since $\lambda_{\min}w^Tw \le w^TRw \le \lambda_{\max}w^Tw$, then

$$w^Tw[\lambda_{\min} - \|w\|_\infty] \le \frac{dE}{dt} \le w^Tw[\lambda_{\max} - \|w\|_\infty] \tag{16}$$

So, if $w(0)$ is initialized as a random vector such that $\|w(0)\|_\infty > \lambda_{\max}$, then we know by equation(16) that

$$\left.\frac{dE(t)}{dt}\right|_{t=0} < 0 \tag{17}$$

So from the continuity of $\frac{dE(t)}{dt}$ we have $\|w(t)\| \le \|w(0)\|$ for all $t \ge 0$. Otherwise, when $\|w(0)\|_\infty \le \lambda_{\max}$, obviously $x(t)$ will never get to a point ξ such that $\|\xi\|_\infty > \lambda_{\max}$, which means that $\|w(t)\|_\infty \le \lambda_{\max}$ for all $t \ge 0$. Then, from the fact $\|w(t)\| \le \sqrt{n}\|w(0)\|_\infty$, it can be concluded that

$$\|w(t)\| \le \max\{\|w(0)\|, \sqrt{n}\lambda_{\max}\}. \tag{18}$$

Similarly, we have

$$\|w(0)\|_\infty < \lambda_{\min} \Rightarrow \|w(t)\| > \|w(0)\| \tag{19}$$

and

$$\|w(0)\| \geq \lambda_{\min} \Rightarrow \|w(t)\|_\infty \geq \lambda_{\min} \Rightarrow \|w(t)\| \geq \lambda_{\min}. \tag{20}$$

Therefore, we can conclude that

$$\|w(t)\| \geq \min\{\|w(0)\|, \lambda_{\min}\} \tag{21}$$

This completes the proof. ■

Theorem 3: If $\xi \neq 0$ is an equilibrium point of the differential equation (10), then ξ is an eigenvector associated with the eigenvalue $\|\xi\|_\infty$. Conversely, if w^* is an eigenvector corresponding to an eigenvalue of symmetric matrix R, there exists a nonzero number α such that αw^* is an equilibrium point of the differential equation (10).

Proof: The proof is straightforward, thus is omitted here. ■

Recurring to the properties of real symmetric matrices, we know $S_1, \cdots, S_n$ construct a normalized perpendicular basis of R^n. So, $w(t) \in R^n$ can be expressed as

$$w(t) = \sum_{i=1}^{n} z_i(t) S_i \tag{22}$$

Theorem 4: If the initial values of the weight vector satisfy $z_1(0) = S_1^T w(0) \neq 0$, then $\lim_{t\to\infty} \frac{w(t)}{z_1(t)} = S_1$, which is an eigenvector corresponding to λ_1.

Proof: Substituting (22) into the formula (10), we get

$$\frac{d}{dt} z_i(t) = \lambda_i z_i(t) - \|w(t)\|_\infty z_i(t), (i = 1, \cdots, n). \tag{23}$$

Denote $\|w\|_\infty$ by $\sigma(t)$, we have

$$z_i(t) = z_i(0) \exp\left(\lambda_i t - \int_0^t \sigma(s) ds\right), (i = 1, \cdots, n). \tag{24}$$

Obviously, $z_i(t) \neq 0$ $(t > 0)$ only if $z_i(0) \neq 0$. For nonzero initial value $x(0) = \sum_{i=1}^n z_i(0) S_i$, denote $r = \min\{i | z_i(0) \neq 0, 1 \leq i \leq n\}$. It is easy to obtain that

$$\frac{z_i(t)}{z_r(t)} = \frac{z_i(0)}{z_r(0)} \exp[(\lambda_i - \lambda_r)t] \tag{25}$$

for $t \geq 0$. Clearly, if $\lambda_i < \lambda_r (i = r+1, \cdots, n)$, we have

$$\lim_{t\to\infty} \frac{z_i(t)}{z_r(t)} = 0 \tag{26}$$

So, under the assumption that $z_1(0) \neq 0$, we can conclude that

$$\lim_{t\to\infty} \frac{z_i(t)}{z_1(t)} = 0, \quad (i = 2, \cdots, n). \tag{27}$$

By theorem 2, we have $z_1(t)$ is bounded. So we have

$$\begin{aligned} \lim_{t\to\infty} \frac{w(t)}{z_1(t)} &= \lim_{t\to\infty} \sum_{i=1}^{n} \frac{z_i(t)}{z_1(t)} S_i \\ &= S_1 + \sum_{i=2}^{n} \frac{z_i(t)}{z_1(t)} S_i \\ &= S_1 \end{aligned} \tag{28}$$

which means that $w(t)$ tends to be an eigenvector in eigenspace associated with λ_1. This completes the proof. ■

III. Numerical experiment

In the following, we will provide a simulation result to illustrate the performance of our proposed neural network model. A data set $D_x = \{(x_i, y_i), i = 1, \cdots, 500\}$ comes from the Gaussian distribution with correlation matrix

$$R = \begin{bmatrix} 10 & 9 \\ 9 & 10 \end{bmatrix}.$$

To show the robustness of the norm based algorithms, we have randomly generated some noise with large magnitude as shown in Fig. 1(a). Thus the unit principal eigenvector $S_1 = [0.70710, 0.70710]^T$ with eigenvalue $\lambda_1 = 10.9$. We simply choose the learning rate $\eta(t) = \frac{0.05}{t}$ which satisfies conditions specified by the theory of stochastic approximation [1][2].

In order to compare the performances, we compute estimated eigenvalue for each learning algorithm at each iteration step by $\tilde{\lambda} = y^2(t)$ for the OJA model (1). Similarly, λ_1 is estimated by $\tilde{\lambda} = \frac{y^2(t)}{w^T(t)w(t)}$ for both the OJAN and LUO algorithms. For the three norm based algorithms, $\tilde{\lambda}$ is given by $\|w(t)\|_1, \|w(t)\|_2^2$ and $\|w(t)\|_\infty$ respectively. After 5000 iterations, we get the estimated principal direction $\xi = [0.69739, 0.71669]^T$ by the proposed infinity norm based PCA algorithm (11) which is shown in Fig. 1(b). Therefore, the proposed algorithm can successfully extract the principal direction of the dataset. As the correlation matrix R is not known beforehand in practice, we thus compute the real-time estimation of λ_1 by $\tilde{\lambda}$ specified above. The error performances $\varepsilon(t) = \tilde{\lambda}(t) - \lambda_1$ for the OJA type algorithms and the norm based algorithms are shown in Fig. 2(a) and Fig. 2(b), respectively. Obviously, the OJA type algorithms behave more oscillations than the norm based algorithms for this noisy dataset. Furthermore, as is shown in Fig. 2(a), unlike the LUO algorithm, the OJA and OJAN algorithms behave nearly the same after some iterations, which means that they are essentially of the same type for both the ODE and the adaptive algorithm. In the case of norm based algorithms, as is shown in Fig. 2(b), the infinity norm based algorithm behaves similar to the $1-$norm based algorithm. However, the proposed infinity norm based algorithm gives the best estimation of λ_1 among the three algorithms. Then, we reach the conclusion that the proposed infinity norm based algorithm demonstrates fast convergence and robustness. Meanwhile, when all of these algorithms are implemented by electronic devices, infinity norm based algorithm (10) may be a better choice due to its simplicity of implementation.

IV. Conclusion

This paper studies two types of learning algorithms for PCA, i.e. the OJA type and the norm based algorithms. Inspired by the existing 1-norm and 2-norm based PCA algorithms, a simple infinity norm based algorithm is proposed. Theoretical analysis shows that it is qualified to extract the principal

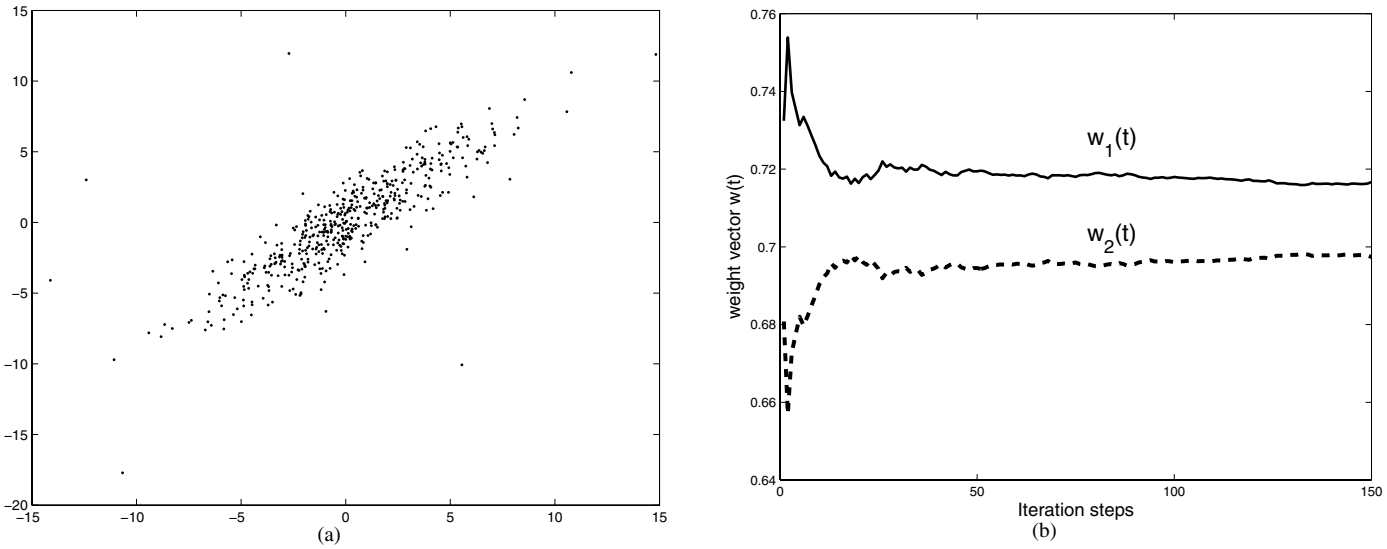

Fig. 1. Experiment result of the infinity norm based PCA algorithm: (a) Dataset drawn from Gaussian data with noise of large magnitude; (b) Asymptotical tendency of $w(t) = [w_1(t), w_2(t)]^T$ to principal direction $[1, 1]^T$.

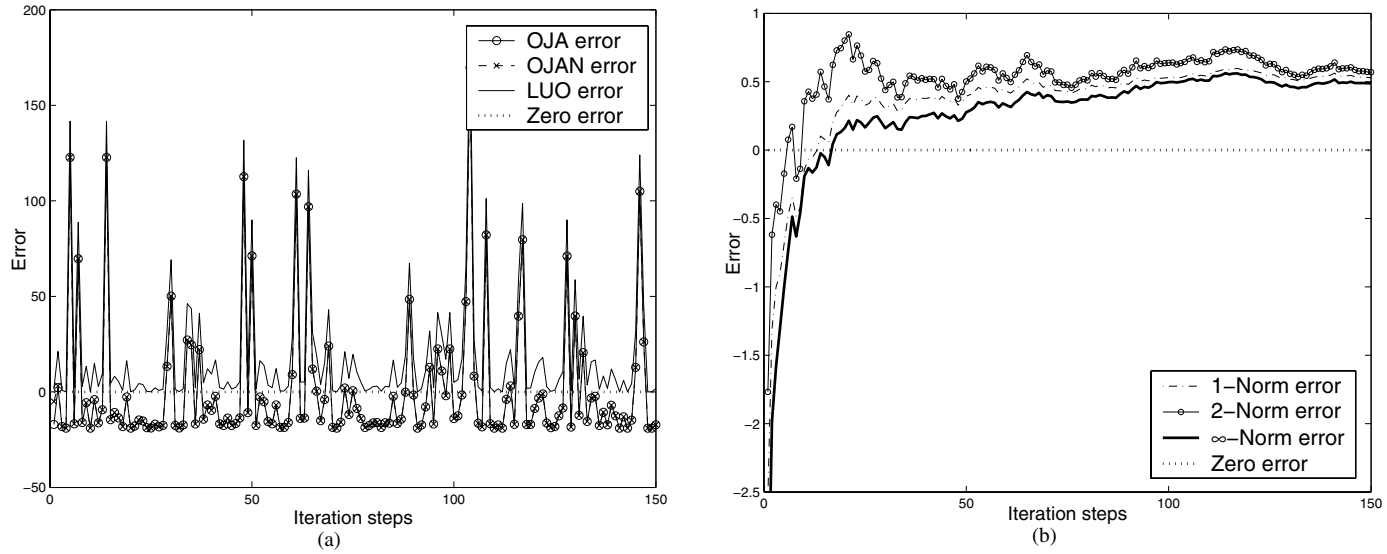

Fig. 2. Error performance $\varepsilon(t) = \tilde{\lambda}(t) - \lambda_1$ of the two different types of algorithms: (a) $\varepsilon(t)$ computed by OJA type algorithms; (b) $\varepsilon(t)$ computed by norm based algorithms;

component. Numerical simulation further shows its fast convergence and robustness compared to the OJA type algorithms. However, it should be noticed that the error $\varepsilon(t) = \tilde{\lambda}(t) - \lambda_1$ is still unsatisfactory even after $w(t)$ reaches its equilibrium as shown in Fig 2.(b). So, further studies should be focused on improving the performance of the proposed algorithm, while maintaining its robustness.

REFERENCES

[1] Z. K. Nie and Z. B. Xu, *Stochastic approximation and adptive algorithms*, Science Press, Beijing, 2003.

[2] L. Ljung, "Analysis of recursive stochastic algorithms", *IEEE Trans. on Auto. Cont.*, vol. 22, pp. 551–575, 1977.

[3] E. Oja, "Principal components, minor components, and Linear neural networks", *Neural Networks*, vol. 5, pp. 927–935, 1922.

[4] Y. Zhang, F. Yan, and H. J. Tang, "Neural networks based approach for computing eigenvectors and eigenvalues of symmetric matrix", *Comp. and Math. with Appl.*, vol. 47, pp. 1155–1164, 2004.

[5] Y. G. Liu, Z. You, and L. Cao, "A simple functional neural network for computing the largest and smallest eigenvalues and corresponding eigenvectors of a real symmetric matrix", *Neurocomputing*, vol. 67, pp. 369–383, 2005.

[6] Y. G. Liu and Z. You, "A concise functional neural network for computing the extremum eigenpairs of real symmetric mtrics", ISNN2006, LNCS 3971, pp. 405-413, 2006.

[7] F. L. Luo, R. Unbehauen, and Y. D. Li, "A principal component analysis algorithm with invariant norm", *Neurocomputing*, pp. 213–221, 1995.

[8] L. Xu, E. Oja, and C. Suen, "Modified hebbian learning for curve and surface fitting", *Neural Networks*, vol. 5, pp. 441–457, 1992.

[9] T. P. Chen and S. Amari, "Unified stabilization approach to principal and minor componets extraction algorithms", *Neural Networks*, vol. 14, pp. 1377–1387, 2001.

[10] Q. F. Zhang and Y. W. Leung, "A class of learning algorithms for

principal component analysis and minor compoent analysis", *IEEE Trans. on Neural Networks*, vol. 11, pp. 200–204, 2000.
[11] R. K. Miller and A. N. Michel, *Ordinary Differential Equations*, New York Academic, 1982.
[12] G. H. Gloub, and C. F. VanLoan, *Matrix Computations*, Johns Hopkins Univ. Press, 1983.