Neural Networks 36 (2012) 64–72

Contents lists available at SciVerse ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Convergence analysis of deterministic discrete time system of a unified self-stabilizing algorithm for PCA and MCA

Xiangyu Kong [a,*], Qiusheng An [b], Hongguang Ma [a], Chongzhao Han [c], Qi Zhang [a]

[a] The Xi'an Research Institute of High Technology, Xi'an Shaanxi 710025, PR China

[b] School of Mathematics and Computer Science, Shanxi Normal University, Linfen Shanxi, 041004, PR China

[c] School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an Shaanxi 710049, PR China

ARTICLE INFO

Article history:
Received 26 December 2011
Revised and accepted 28 August 2012

Keywords:
Deterministic discrete-time (DDT) system
Feature extraction
Minor component analysis (MCA)
Neural networks
Principal component analysis (PCA)

ABSTRACT

Unified algorithms for principal and minor components analysis can be used to extract principal components and if altered simply by the sign, it can also serve as a minor component extractor. Obviously, the convergence of these algorithms is an essential issue in practical applications. This paper studies the convergence of a unified PCA and MCA algorithm via a corresponding deterministic discrete-time (DDT) system and some sufficient conditions to guarantee convergence are obtained. Simulations are carried out to further illustrate the theoretical results achieved.

1. Introduction

Principal component analysis (PCA) and minor component analysis (MCA) provide powerful techniques in many information processing fields. For example, PCA is a useful tool in feature extraction, data compression, pattern recognition, time series prediction, etc. (Lv, Zhang & Tan, 2006), and MCA has been applied in frequency estimation, bearing estimation, digit beamforming, moving target indication, clutter cancellation, total least squares, computer vision, etc (Cirrincione, Cirrincione, Herault, & Huffel, 2002). Neural networks can be used to solve the task of PCA and MCA, which possess many obvious advantages, such as lower computational complexity and better suitability for high-dimensional and nonstationary data, compared with the traditional algebraic approaches. In the last few decades, many neural network learning algorithms were proposed to extract principal components (Bannour & Azimi-Sadjadi, 1995; Cichocki, Kasprzak, & Skarbek, 1996; Kung, Diamantaras, & Taur, 1994; Möller & Könies, 2004; Oja, 1982; Ouyang, Bao, & Liao, 2000; Sanger, 1989; Xu, 1993; Yu, Efeand, & Kaynak, 2002; Yu, Poznyak, & Li, 2001) or minor components (Cirrincione et al., 2002; Douglas, Kung, & Amari, 2002; Feng, Bao, & Jiao, 1998; Feng, Zheng, & Jia, 2005; Kong, Hu, & Han, 2010b; Luo & Unbehauen, 1997; Möller, 2004; Oja, 1992; Ouyang, Bao, Liao, & Ching, 2001; Xu, Oja, & Suen, 1992; Zhang & Leung, 2000), respectively. Clearly, a unified neural network algorithm, capable of both PCA and MCA by simply switching the sign in the same learning rule, is of practical significance in the implementations of algorithms, which can reduce the complexity and cost of hardware implementations (Chen & Amari, 2001; Hasan, 2007). In this research area, many pioneering works have been proposed by Chen and Amari (2001). Recently, a few self-normalizing dual systems for minor and principal component extraction are proposed and their stability is widely analyzed (Hasan, 2007; Kong, Hu, & Han, 2012; Peng, Zhang, & Xiang, 2009).

The convergence of neural network learning algorithms is a difficult topic for direct study and analysis. Traditionally, based on the stochastic approximation theorem, the convergence of these neural network learning algorithms is interpreted indirectly by analyzing corresponding deterministic continuous time (DCT) systems (Kushner & Clark, 1976; Ljung, 1977). The stochastic approximation theorem requires that some restrictive conditions must be satisfied. One important condition is that the learning rates of algorithms must approach zero. Clearly, the restrictive condition is difficult to satisfy in many practical applications, where a constant learning rate is usually employed due to computational roundoff limitations and tracking requirements (Zufiria, 2002). Recently, deterministic discrete time (DDT) systems have been proposed instead to indirectly interpret the dynamics of the neural network learning algorithms described by stochastic discrete time (SDT) systems (Zufiria, 2002), and many MCA or PCA algorithms are analyzed via the DDT method (Kong, Hu, & Han, 2010a; Lv et al., 2006;

* Corresponding author.
E-mail address: xiangyukong01@163.com (X. Kong).

0893-6080/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2012.08.016

Mao, Fan, & Li, 2006; Peng, Zhang & Xiang, 2008; Zhang, 2003; Zhang, Mao, Lv, & Tan, 2005; Zufiria, 2002). It is worth noting that DDT systems allow learning rates to be constant and can preserve the discrete time nature of the original SDT systems.

Despite the large number of unified PCA and MCA algorithms proposed to date, there are few papers that analyze these algorithms via DDT methods and derive the conditions to guarantee convergence. Obviously, it is necessary to perform such research for these algorithms from the point of application. Among the algorithms of the unified PCA and MCA, Chen's algorithm (Chen & Amari, 2001) is regarded as pioneering work. Other self-normalizing dual systems (Hasan, 2007) or dual purpose algorithms (Kong et al., 2012; Peng et al., 2009) can be viewed as the generalizations of the unified Chen algorithm (Chen & Amari, 2001). Chen's algorithm lays sound theoretical foundations for the dual purpose algorithm research. However, no work has been done so far on the study of Chen's DDT system. In this paper, the unified PCA and MCA algorithm proposed by Chen and Amari (2001) is analyzed and the sufficient conditions to guarantee convergence are derived by the DDT method, and these theoretical results can lay a solid foundation for the application of this algorithm.

This paper is organized as follows. In Section 2, a unified algorithm for PCA and MCA is presented. In Section 3, the convergence analysis of the unified algorithm for PCA and MCA via the DDT method is given. In Section 4, computer simulation results on extracting the principal component and minor component are presented. Finally, we give some conclusions in Section 5.

2. Unified self-stabilizing algorithm for PCA and MCA

Consider a single linear neuron with the following input–output relation: $y(k) = \boldsymbol{W}^T(k)\boldsymbol{X}(k)$, $k = 0, 1, 2, \ldots$, where $y(k)$ is the neuron output, the input sequence $\{\boldsymbol{X}(k)|\boldsymbol{X}(k) \in \mathbb{R}^n(k = 0, 1, 2, \ldots)\}$ is a zero mean stationary stochastic process, and $\boldsymbol{W}(k) \in \mathbb{R}^n(k = 0, 1, 2, \ldots)$ is the weight vector of the neuron. In the last few decades, many neural network learning algorithms (Bannour & Azimi-Sadjadi, 1995; Cichocki et al., 1996; Kung et al., 1994; Möller & Könies, 2004; Oja, 1982) have been proposed to update the weight vector $\boldsymbol{W}(k)$, such that as $k \to \infty$, $\boldsymbol{W}(k)$ can converge to principal components or minor components, respectively. Chen and Amari (2001) proposed a unified stabilizing learning algorithm for principal components and minor components extraction, and the stochastic discrete form of the algorithm can be written as

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) \pm \eta[\|\boldsymbol{W}(k)\|^2 y(k)\boldsymbol{X}(k) - y^2(k)\boldsymbol{W}(k)] + \eta(1 - \|\boldsymbol{W}(k)\|^2)\boldsymbol{W}(k), \quad (1)$$

where $\eta(0 < \eta < 1)$ is the learning rate. The algorithm (1) can extract principal components if "+" is used. If the sign is simply altered, (1) can also serve as a minor component extractor. It is interesting that the only difference between the PCA algorithm and MCA algorithm is the sign on the right hand side of (1). This is of practical significance in neural networks implementations.

In order to derive the sufficient conditions to guarantee convergence of algorithm (1), next, we analyze the dynamics of (1) via the DDT approach. The DDT system associated with (1) can be formulated as follows. Taking the conditional expectation $E\{\boldsymbol{W}(k+1)/\boldsymbol{W}(0), \boldsymbol{X}(i), i < k\}$ operator to (1) and identifying the conditional expected value as the next iterate, a DDT system can be obtained and given as

$$\boldsymbol{W}(k+1) = \boldsymbol{W}(k) \pm \eta \times [\|\boldsymbol{W}(k)\|^2 \boldsymbol{R}\boldsymbol{W}(k) - \boldsymbol{W}^T(k)\boldsymbol{R}\boldsymbol{W}(k)\boldsymbol{W}(k)] + \eta(1 - \|\boldsymbol{W}(k)\|^2)\boldsymbol{W}(k), \quad (2)$$

where $\boldsymbol{R} = E[\boldsymbol{X}(k)\boldsymbol{X}^T(k)]$ is the correlation matrix of input data. The main purpose of this paper is to study the convergence characteristics of the weight vector $\boldsymbol{W}(k)$ of (2) subject to the learning rate η being some constant. It is worth mentioning that Chen's algorithm has a computational complexity $2Nr^2 + 3Nr$, compared with $O(N^3)$ of the inverse-power iteration algorithm and $O(Nr^2)$ of the Rayleigh quotient iteration. We can see that the computational complexity of Chen's algorithm is of the same order as that of the Rayleigh quotient iteration and is smaller than that of the inverse-power iteration algorithm. The strong point of Chen's algorithm is that it is a unified neural networks algorithm, capable of both PCA and MCA by simply switching the sign in the same learning rule.

3. Convergence analysis

For convenience of analysis, some preliminaries are given. Since $\boldsymbol{R}$ is a symmetric positive definite matrix, there exists an orthonormal basis of $\mathbb{R}^n$ composed of the eigenvectors of $\boldsymbol{R}$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ to be all the eigenvalues of $\boldsymbol{R}$ ordered by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} \geq \lambda_n > 0$. Denote by σ, the largest eigenvalue of $\boldsymbol{R}$. Suppose that the multiplicity of σ is $m(1 \leq m \leq n)$, then $\sigma = \lambda_1 = \cdots = \lambda_m$. Suppose that $\{\boldsymbol{V}_i|i = 1, 2, \ldots, n\}$ is an orthogonal basis of $\mathbb{R}^n$ such that each $\boldsymbol{V}_i$ is a unit eigenvector of $\boldsymbol{R}$ associated with the eigenvalueλ_i. Denote by $\boldsymbol{V}_\sigma$ the eigensubspace of the largest eigenvalueσ, i.e. $\boldsymbol{V}_\sigma = span\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m\}$. Denoting by $\boldsymbol{V}_\sigma^\perp$ the subspace which is perpendicular to $\boldsymbol{V}_\sigma$, clearly $\boldsymbol{V}_\sigma^\perp = span\{\boldsymbol{V}_{m+1}, \ldots, \boldsymbol{V}_n\}$. Similarly, we can denote by $\boldsymbol{V}_\tau$ eigensubspace of the smallest eigenvalue τ. Suppose that the multiplicity of τ is $p(1 \leq p \leq n - m)$, then $\boldsymbol{V}_\tau = span\{\boldsymbol{V}_{n-p}, \ldots, \boldsymbol{V}_n\}$ and $\boldsymbol{V}_\tau^\perp = span\{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{n-p-1}\}$.

Since the vector set $\{\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_n\}$ is an orthonormal basis of $\mathbb{R}^n$, for each $k \geq 0$, $\boldsymbol{W}(k)$ and $\boldsymbol{R}\boldsymbol{W}(k)$ can be represented respectively as

$$\boldsymbol{W}(k) = \sum_{i=1}^{n} z_i(k)\boldsymbol{V}_i, \qquad \boldsymbol{R}\boldsymbol{W}(k) = \sum_{i=1}^{n} \lambda_i z_i(k)\boldsymbol{V}_i, \quad (3)$$

where $z_i(k)(i = 1, 2, \ldots, n)$ is some constant.

From (2) and (3), it holds that

$$z_i(k+1) = [1 \pm \eta(\lambda_i\|\boldsymbol{W}(k)\|^2 - \boldsymbol{W}^T(k)\boldsymbol{R}\boldsymbol{W}(k)) + \eta(1 - \|\boldsymbol{W}(k)\|^2)]z_i(k), \quad (4)$$

$(i = 1, 2, \ldots, n)$, for all $k \geq 0$. By denoting the function $Q(\boldsymbol{R}, \boldsymbol{W}(k)) = \pm[\lambda_i\|\boldsymbol{W}(k)\|^2 - \boldsymbol{W}^T(k)\boldsymbol{R}\boldsymbol{W}(k)]$, (4) can be represented as

$$z_i(k+1) = [1 + \eta Q(\boldsymbol{R}, \boldsymbol{W}(k)) + \eta(1 - \|\boldsymbol{W}(k)\|^2)]z_i(k), \quad (5)$$

$(i = 1, 2, \ldots, n)$, for all $k \geq 0$. According to the relevant properties of the Rayleigh Quotient (Cirrincione et al., 2002), it clearly holds that

$$\lambda_n\boldsymbol{W}^T(k)\boldsymbol{W}(k) \leq \boldsymbol{W}^T(k)\boldsymbol{R}\boldsymbol{W}(k) \leq \lambda_1\boldsymbol{W}^T(k)\boldsymbol{W}(k), \quad (6)$$

for all $k \geq 0$. From (6), it holds that

$$Q_{\max} = (\lambda_1 - \lambda_n)\|\boldsymbol{W}(k)\|^2, \qquad Q_{\min} = (\lambda_n - \lambda_1)\|\boldsymbol{W}(k)\|^2. \quad (7)$$

Next, we analyze the convergence of DDT system (2) via the following Theorems 1–4. Proofs of Theorems 1–4 are given in the Appendix.

Theorem 1. *Suppose that $\eta \leq 0.3$, if $\|\boldsymbol{W}(0)\| \leq 1$ and $(\lambda_1 - \lambda_n) < 1$, then it holds that $\|\boldsymbol{W}(k)\| < (1 + \eta\lambda_1)$, for all $k \geq 0$.*

Theorem 1 shows that there exists an upper bound of $\|\boldsymbol{W}(k)\|$ in the DDT system (2), for all $k \geq 0$.