Convergence Analysis of Online Linear Discriminant Analysis

K. Hiraoka, S. Yoshizawa, K. Hidai, M. Hamahira,
H. Mizoguchi, and T. Mishima
Dept. of Information and Computer Sciences, Saitama University
255 Shimo-okubo, Urawa, 338-8570, Japan
email: hira@ics.saitama-u.ac.jp

Abstract

Convergence of a matrix dynamics for online LDA is analyzed. Especially, stable spurious solutions are pointed out and two schemes to prevent the spurious solutions are proposed. The performance of the algorithm is confirmed by simulations of face identification.

1 Introduction

Linear discriminant analysis (LDA) is applied to broad areas, e.g. image recognition [5]. However, online algorithms of LDA are not sufficiently studied while online principal component analysis (PCA) has been established well [1][2]. Note that $N \times N$ matrices appears in [3][4], where N is the dimension of data. This weak point is serious because N is often large for some applications such as image recognition tasks. Recently, an online LDA algorithm which does not need $N \times N$ matrices is proposed [6][7]. In the present paper, convergence of this algorithm is analyzed.

2 Online LDA Algorithm

To estimate the class for a data vector $\boldsymbol{x}$ by LDA, we transform it to the "feature" vector $\boldsymbol{y} = A^T\boldsymbol{x}$ and compare $\boldsymbol{y}$ with the mean feature vector of each class. In online LDA, the discriminant matrix A is updated every time new data is presented.

2.1 Matrix dynamical system for LDA

Let "between-class" covariance B and "within-class" covariance W be $N \times N$ symmetric matrices. In this paper, B and W are assumed to be positive semidefinite and positive definite, respectively. When a scalar λ and a vector $\boldsymbol{p} \neq \boldsymbol{0}$ satisfy $B\boldsymbol{p} = \lambda W\boldsymbol{p}$, λ and $\boldsymbol{p}$ are called a *generalized eigenvalue* and a *generalized eigenvector*, respectively.

In LDA (esp. Fisher Linear Discriminant), we have to find L generalized eigenvectors $\boldsymbol{p}_1, \cdots, \boldsymbol{p}_L$ which correspond to L largest generalized eigenvalues and $\|\boldsymbol{p}_1\| = \cdots = \|\boldsymbol{p}_L\| = 1$, where $L \leq N$ is the number of "features". To be exact, it is enough to find an $N \times L$ matrix $A = (a_{ij})$ which can be written as $A = P_L Q$, where Q is an arbitrary orthogonal matrix and $P_L = (\boldsymbol{p}_1, \cdots, \boldsymbol{p}_L)$. Let $\mathcal{A}_L$ be the set of all such As. Note that $A \in \mathcal{A}_L$ satisfies $BA = WA\Gamma$ and $A^TWA = I$, where Γ is an $L \times L$ matrix and I is the identity matrix.

In order to find a matrix $A \in \mathcal{A}_L$, a potential function is considered:

$$\phi(A) = \mathrm{Tr}\left[A^T B A \left(I - \frac{1}{2} A^T W A\right)\right]. \tag{1}$$

This ϕ takes its maximum value on $\mathcal{A}_L$ [7]. By differentiating ϕ, we obtain a potential flow $da_{ij}/dt = (1/2)(\partial\phi(A)/\partial a_{ij})$. This flow is calculated as

$$\frac{d}{dt}A(t) = BA(t) - \frac{1}{2}BA(t)A(t)^T WA(t) - \frac{1}{2}WA(t)A(t)^T BA(t). \tag{2}$$

From this matrix dynamical system (2), the following *online LDA* algorithm is constructed.

2.2 Online LDA algorithm

At every time step $t = 1, 2, 3, \cdots$, a new pair $(\boldsymbol{x}(t), c(t))$ is presented, where $\boldsymbol{x}(t)$ is a data vector, $c(t) \in \{1, \cdots, M\}$ is the class of $\boldsymbol{x}(t)$, and M is the number of classes. The number L of features is less than M. Based on this pair, auxiliary variables are updated as follows:

$$t^c(t) = t^c(t-1) + \delta(c, c(t)), \quad \bar{\boldsymbol{x}}(t) = \left(1 - \frac{1}{t}\right)\bar{\boldsymbol{x}}(t-1) + \frac{1}{t}\boldsymbol{x}(t), \tag{3}$$

$$\bar{\boldsymbol{x}}^c(t) = \begin{cases} \left(1 - \frac{1}{t^c(t)}\right)\bar{\boldsymbol{x}}^c(t-1) + \frac{1}{t^c(t)}\boldsymbol{x}(t) & (c = c(t)), \\ \bar{\boldsymbol{x}}^c(t-1) & (c \neq c(t)), \end{cases} \tag{4}$$

$$\boldsymbol{v}^c(t) = \bar{\boldsymbol{x}}^c(t) - \bar{\boldsymbol{x}}(t), \quad \boldsymbol{w}(t) = \boldsymbol{x}(t) - \bar{\boldsymbol{x}}^{c(t)}(t), \tag{5}$$

$$\boldsymbol{y}^c(t) = A(t-1)^T \boldsymbol{v}^c(t), \quad \boldsymbol{z}(t) = A(t-1)^T \boldsymbol{w}(t), \tag{6}$$

$$F(t) = \frac{1}{M}\sum_{c=1}^{M} \boldsymbol{v}^c(t)\boldsymbol{y}^c(t)^T, \quad \boldsymbol{g}(t) = \frac{1}{M}\sum_{c=1}^{M} \boldsymbol{y}^c(t)\left(\boldsymbol{y}^c(t)^T \boldsymbol{z}(t)\right), \tag{7}$$

where $c = 1, \cdots, M$ and $\delta(c, c(t)) = 1\,(c = c(t)),\ 0\,(c \neq c(t))$. Then the discriminant matrix A is updated as

$$A(t) = A(t-1) + \eta\Big(F(t) - \frac{1}{2}F(t)\boldsymbol{z}(t)\boldsymbol{z}(t)^T - \frac{1}{2}\boldsymbol{w}(t)\boldsymbol{g}(t)^T\Big), \tag{8}$$

where $\eta > 0$ is the learning coefficient. Note that the variables $\boldsymbol{y}^c, \boldsymbol{z}, F, \boldsymbol{g}$ are introduced instead of calculating $B(t) = \frac{1}{M}\sum_{c=1}^{M} \boldsymbol{v}^c(t)\boldsymbol{v}^c(t)^T$ itself so that $N \times N$ matrices are not needed to update A. In addition, $W(t) = \frac{1}{t}\sum_{\tau=1}^{t} \boldsymbol{w}(\tau)\boldsymbol{w}(\tau)^T$ is replaced with the instantaneous value $\boldsymbol{w}(t)\boldsymbol{w}(t)^T$.

As for the initial values, $t^c(0) = 0$, $\bar{\boldsymbol{x}}(0)$ and $\bar{\boldsymbol{x}}^c(0)$ are arbitrary vectors, and $A(0)$ is an arbitrary matrix which satisfies $\mathrm{rank}A(0) = L$.

3 Fixed Points and Their Stability

Let $J_B = A^T B A$ and $J_W = A^T W A$.

Theorem 1 *The following conditions are equivalent:*

1. *A is a fixed point of (2).*
2. *A satisfies*

$$BA = WAJ_B, \tag{9}$$

$$J_B = J_W J_B = J_B J_W. \tag{10}$$

Corollary 1 *Let A be a fixed point of (2). Then, $J_W = I$ unless J_B is singular. In other words, J_B has an eigenvalue 0 if $J_W \neq 0$.*

Corollary 2 *Let A be a fixed point of (2). Then, by taking an appropriate orthogonal matrix R and replacing AR by A, we obtain*

$$J_B = \mathrm{diag}(\beta_1, \cdots, \beta_K, 0, \cdots, 0), \tag{11}$$

$$J_W = \mathrm{diag}(1, \cdots, 1, \omega_{K+1}, \cdots, \omega_L), \tag{12}$$

where $\beta_1 \geq \cdots \geq \beta_K > 0$ and $\omega_{K+1} \geq \cdots \geq \omega_L$.

Theorem 2 *Let A be a fixed point of (2) and $J_W = I$. Then this fixed point A is stable if and only if $A \in \mathcal{A}_L$.*

Theorem 3 *Assume that* $\mathrm{rank}B \geq L$. *Let A be a fixed point of (2) and $J_W \neq I$. Then this fixed point A is stable if and only if the following conditions hold in Corollary 2:*

1. *$\omega_{K+1}, \cdots, \omega_L > 2$, and*
2. *$\beta_1, \cdots, \beta_K$ are L largest generalized eigenvectors.*

4 Preventing Spurious Solutions

Theorem 2 corresponds to *true* solutions while Theorem 3 corresponds to *spurious* solutions. In this section, we discuss simple schemes to prevent the spurious solutions.

4.1 Starting around the origin

From Theorem 2 and Theorem 3, $\mathrm{Tr}(A^T W A) = L$ at true solutions, while $\mathrm{Tr}(A^T W A) > L+1$ at spurious solutions. In this sense, spurious solutions are more "far" from the origin than true solutions. This fact leads us to a heuristic scheme: take the initial value of A near the origin.

This scheme works well in our simulations. It also has a theoretical ground at least for the case $L = 1$, namely, A is a vector.

Theorem 4 *Assume that $L = 1$ and $A(0)^T W A(0) < 2$. Then the solution of (2) keeps $A(t)^T W A(t) < 2$ for all $t > 0$.*

This theorem is proved since

$$\begin{aligned} &\frac{d}{dt}\left(A(t)^T W A(t)\right) \\ &= 2A(t)^T W B A(t)\left(1 - \frac{1}{2}A(t)^T W A(t)\right) - A(t)^T W^2 A(t) A(t)^T B A(t) \end{aligned} \tag{13}$$

is negative or zero when $A(t)^T W A(t) = 2$. Remember that there is no stable fixed point in the region $A^T W A \leq 2$ except for the true solutions $A \in \mathcal{A}_L$. Moreover, there is no periodical orbit because (2) is a potential flow.

4.2 Relaxation of B

Another simple scheme is replacing B with $B + \epsilon_B I$, where ϵ_B is a small positive number. Then this *new* B is positive definite now. This means that all fixed points are unstable except for the true solutions because of the following reason. Assume that A is a fixed point of (2), B is positive definite, and $J_W \neq I$. From Corollary 1, J_B has an eigenvalue 0. Then rank$A$ must be less than L since B is positive definite. In this case, $\omega_L = 0 < 2$ and this A is unstable from Theorem 3.

If one does not allow the small error of the solution which is caused by $+\epsilon_B I$, one can use $+\epsilon_B W$ instead. Then the generalized eigenvectors does not change.

5 When Sample Size Is Small

When the number t of samples is small compared with the dimension L of data, W can be singular and $A(t)$ can diverge. In this case, replacing W with $W + \epsilon_W I$ prevents the divergence of $A(t)$, where ϵ_W is a small positive number. This replacement is also desirable in terms of robustness because a component which correspond to a too small eigenvalue of W is useless noise in most cases.

6 Simulation

We have applied the proposed algorithm for face identification task. The face image consists of 10×10 pixels, and hence $N = 100$. Each pixel has 256 level gray scale value from -1 to $+1$. The number of persons is $M = 3$ in the first part of the simulation. In the latter part, new person is added and $M = 4$. For each person, 100 images are used as the sample for this experiment. The parameters of learning are $L = 2$, $\eta = 0.01$, $\epsilon_B = 0$, $\epsilon_W = 0.01$. The result is shown in Fig. 1. The step t in the figure is equal to the total number of presented images. Online LDA successfully adapts to the new situation.

We have also examined the effect of the initial value. After sufficient learning ($t = 40000$) with $N = 100$, $M = 3$, $L = 2$, $\eta = 0.001$, $\epsilon_B = 0$ and $\epsilon_W = 0.0001$, only 93% of 100 trials attained 100% correct identification when each element of $A(0)$ is generated by the uniform distribution on $[-1, +1]$, while all 303 trials attained 100% when on $[-0.01, +0.01]$. This result is consistent with the discussion in section 4.1.

7 Conclusion

The convergence of the matrix dynamics for online LDA are analyzed. Especially, all fixed points are identified and their stability is determined. Then two schemes to prevent the spurious solutions are proposed. The performance of the algorithm is confirmed by simulations.

One problem in online LDA is selection of the learning coefficient η. A guideline on selection or automatic adjustment of η is desired to use online LDA for extensive applications.

From another potential function $\psi(A) = \log\det[A^T BA] - \mathrm{Tr}[A^T WA]$ together with the metric $\langle \Delta A_1, \Delta A_2 \rangle_A = \mathrm{Tr}[\Delta A_1^T \Delta A_2 (A^T BA)^{-1}]$, we can obtain another potential flow

$$\frac{d}{dt}A(t) = BA(t) - WA(t)A(t)^T BA(t). \tag{14}$$

As a special case of (14), we obtain Oja's flow [1] when $W = I$. However, convergence of the algorithm based on (14) is significantly slower than the convergence of the present algorithm. The reason of this phenomenon is not clear at now.

This work has been partly supported by CREST of JST (Japan Science and Technology) 279102.

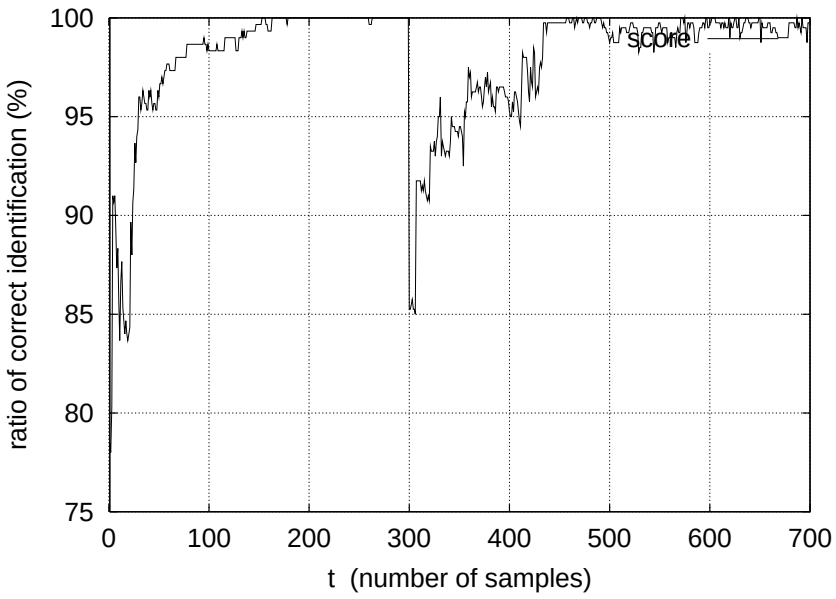

Fig. 1: **Learning curve of online LDA.**
Until $t = 300$, images of three persons are presented. At $t = 300$, a new class of the fourth person is added. After $t = 300$, images of four persons are presented one by one. Online LDA adapts to the new situation.

References

[1] E. Oja, "A simplified neuron model as a principal component analyzer", J. Math. Biol., Vol. 15, pp. 267–273, 1982.

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

[3] J. Mao and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection", IEEE Tr. NN, Vol. 6, No. 2, pp. 296–317, 1995.

[4] C. Chatterjee and V. P. Roychowdhury, "On self-organizing algorithms and networks for class-separability features", IEEE Tr. NN, Vol. 8, No. 3, pp. 663–678, 1997.

[5] T. Kurita and S. Hayamizu, "Gesture Recognition using HLAC Features of PARCOR Images and HMM based Recognizer", Proceedings of the third international conference on automatic face and gesture recognition, pp. 422-427, April, 1998.

[6] K. Hiraoka and M. Hamahira, "On Successive Learning Type Algorithm for Linear Discriminant Analysis", IEICE Technical Report, NC99-73, pp. 85–92, 1999 (in Japanese).

[7] K. Hiraoka, K. Hidai, et al., "Derivation of online LDA algorithm and its application for face identification", Fifth Robotics Symposia, 2000 (in Japanese), *submitted.*