Regular Paper

The Statistical Properties of the Learning Subspace Methods for Pattern Recognition

HUANG DESHUANG†

This paper studies the statistical and disturbed properties for self-supervised learning subspace methods for pattern recognition. We prove that the transformed matrices of the LSM converge to the estimate of pattern auto-correlation matrix, and give the approximate expression of the eigen spectrum caused by the LSM's rotation. Finally the upper bound disturbed theorem for Learning Subspace is proposed and proved.

1. Introduction

The subspace method was introduced by S. Watanabe in 1967[1)], which indicates that the structure information corresponding to each pattern category clusters according to a subspace. Afterwards, some experts presented several new and different methods for determining category subspaces[2)~4)]. In 1978, T. Kohonen proposed a new concept called learning subspace methods (LSM), and developed the idea of the subspace for pattern recognition, which utilizes the sequential input samples to modify or learn (rotate) the formed subspaces so that the principal eigen vectors corresponding to the modified pattern subspaces can trace and express the changing pattern information[4)]. In 1982, E. Oja proposed an **averaging learning subspace method** (ALSM)[5),6)] which can avoid the sensitivity to the order of the input samples, but needs to compute three conditioned correlation matrices and their eigenvalue decomposition resulting in decreasing the convergence speed[6)]. In order to avoid the defects of these methods, the author proposed three kinds of learning subspace methods, i.e., **minimum norm learning subspace method** (MNLSM), **detecting error averaging learning subspace method** (DEALSM), **forward-backward smoothing learning subspace method** (FBSLSM)[7)~9)].

However, how are the steadiness and convergence of all these methods? Can they classify those samples not to be learned, i.e., possess generalization capabilities? These problems have not completely been solved. In fact, the Kohonen's self-organizing Learning Subspace Method (LSM) is a self-supervised learning neural network which self-learns and self-organizes the basis vectors in pattern subspace using input pattern samples[4)].

The generalization capabilities in neural networks are very important aspects for improving the network performances. The abilities to generalize is meant responding properly to previously unseen input data[9)]. In the case of the network used as a pattern classifier, this means classifying correctly samples that have not been used in training the networks. Poor generalization can be caused in two cases: (1) the selected self-supervised method; (2) the quality and quantity of training samples available, which makes a main effect on the robustness and disturbance of the clustering subspace formed. So we study the robustness and disturbance of the self-supervised LSM in this paper.

This paper is organized as follows. The concepts of the LSMs for pattern recognition and their rules for classification are discussed in Section 2. Section 3 presents the distance measure's theorem for subspaces and the orthogonal iterating convergence theorem for subspaces. Section 4 discusses the robustness and disturbance properties of the self-supervised LSM, and gives the convergence solution of transformed matrices of learning subspaces and the approximate expression of the eigen spectrum caused by LSM's rotation, and presents and proves the upper bound disturbed theorem for learning subspace. Finally, several conclusions are given in Section 5.

2. Learning Subspace Method for Pattern Recognition

Assume that the ith one of the K categories $\{\omega_i,\ i = 1, 2, \cdots, K\}$ corresponds to N_i pat-

† National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

tern sample vectors $\{\boldsymbol{x}_k^{(i)},\ i=1,2,\cdots,K, k=1,2,\cdots,N_i\}$. The learning rules of Kohonen self-supervised LSMs are expressed as follows:

$$\begin{cases} L_k^{(i)} = (I + \mu_1 \boldsymbol{x}_k^{(i)} \boldsymbol{x}_k^{(i)T}) L_{k-1}^{(i)} \\ L_k^{(j)} = (I - \mu_2 \boldsymbol{x}_k^{(i)} \boldsymbol{x}_k^{(i)T}) L_{k-1}^{(j)} \\ \qquad (j \neq i = 1,2,\cdots,K) \\ L_k^{(i)} = L[\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, \cdots, \boldsymbol{u}_{p^{(i)}}^{(i)}] \end{cases} \quad (1)$$

where μ_1, μ_2 are learning coefficients which can be related to the learning vector $\boldsymbol{x}_k^{(i)}$. In general, $|\mu_1|$ or $|\mu_2| < 1/||\boldsymbol{x}_k^{(i)}||^2$; T denotes transform of the matrix; $L_k^{(i)}$ indicates the ith subspace composed of $p^{(i)}$ basis vectors $\boldsymbol{u}_n^{(i)}$ ($n = 1,2,\cdots,p^{(i)}$) at instant k.

Assume that the general subspace rotation is:

$$L' = (I + \mu \boldsymbol{x}\boldsymbol{x}^T)L$$
$$L = L(\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p) \quad (2)$$

where μ is the learning coefficient; p is the dimension of subspace L. For an arbitrary basis vector $\boldsymbol{u}_i$ ($1 \le i \le p$) in subspace L, then the basis vector $\boldsymbol{u}_i'$ transformed by Eq. (2) is:

$$\boldsymbol{u}_i' = (I + \mu \boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{u}_i = \boldsymbol{u}_i + \mu \hat{\boldsymbol{x}} \cdot \boldsymbol{x} \quad (1 \le i \le p) \quad (3)$$

where $\hat{\boldsymbol{x}} = \boldsymbol{x}^T \boldsymbol{u}_i$. This result indicates that the basis vector $\boldsymbol{u}_i'$ in new subspace L' is the basis vector $\boldsymbol{u}_i$ in old subspace L plus the scaling term of the $\boldsymbol{x}$, thus the projection of the $\boldsymbol{x}$ on the $\boldsymbol{u}_i'$ of L' becomes[9]:

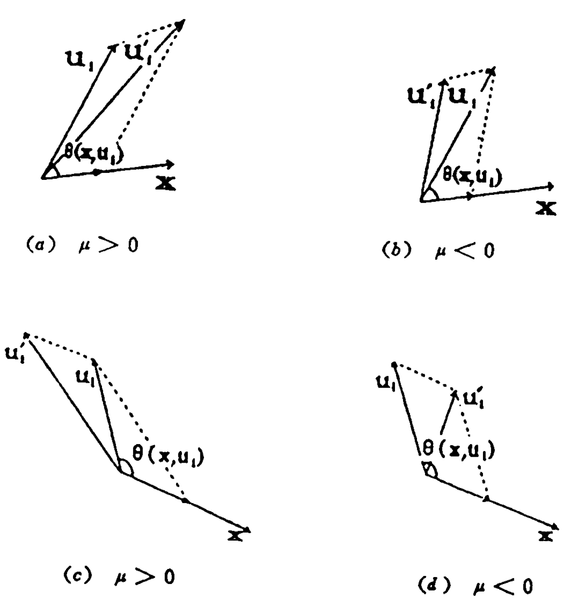

Fig. 1 The geometric explanations for the LSMs.

$$\hat{\boldsymbol{x}}' = \boldsymbol{x}^T \boldsymbol{u}_i + \mu \boldsymbol{x}^T \boldsymbol{u}_i \cdot ||\boldsymbol{x}||^2 = \hat{\boldsymbol{x}}(1 + \mu \cdot ||\boldsymbol{x}||^2) \quad (4)$$

Because of $|\mu| < 1/||\boldsymbol{x}||^2$, so if $\hat{\boldsymbol{x}} > 0$, i.e., $0 \le \theta(\boldsymbol{x}, \boldsymbol{u}_i) \le 90°$ then: (1) $\mu > 0$, $\hat{\boldsymbol{x}}' > \hat{\boldsymbol{x}}$, as is shown in **Fig. 1** (a); (2) $\mu < 0$, $\hat{\boldsymbol{x}}' < \hat{\boldsymbol{x}}$, as is shown in Fig. 1 (b). If $\hat{\boldsymbol{x}} < 0$, i.e., $90° < \theta(\boldsymbol{x}, \boldsymbol{u}_i) \le 180°$, then: (3) $\mu > 0$, $\hat{\boldsymbol{x}}' < \hat{\boldsymbol{x}}$, (or $||\hat{\boldsymbol{x}}'|| > ||\hat{\boldsymbol{x}}||$), as is shown in Fig. 1 (c); (4) $\mu < 0$, $\hat{\boldsymbol{x}}' > \hat{\boldsymbol{x}}$ (or $||\hat{\boldsymbol{x}}'|| < ||\hat{\boldsymbol{x}}||$), as is shown in Fig. 1 (d).

According to above geometric explanations and learning principles for the LSMs, the classifying rule for the Kohonen's learning subspace method is presented as follows:

For an arbitrary pattern vector $\boldsymbol{x}$, if

$$\boldsymbol{x}^T P^{(i)} \boldsymbol{x} > \boldsymbol{x}^T P^{(j)} \boldsymbol{x} \quad \text{for all} \quad j \neq i \quad \text{then} \quad x \in \omega_i \quad (5)$$

where $P^{(i)}$ and $P^{(j)}$ are respectively the orthogonal projection matrices of subspace $L^{(i)}$ and $L^{(j)}$.

3. Distance Measure and Convergence Theorem for the LSMs

3.1 The Angle and Distance between Two LSMs[9]

Assume that $X_1 \in R^{d \times p}$, $Y_1 \in R^{d \times q}$ are the matrices composed of orthogonal normalized basis vectors of the subspace S_1, S_2 in R^d, respectively. $X_1 = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p]$, $Y_1 = [v_1, v_2, \cdots, v_q]$, where supposes $p < q$, $2p \le d$.

We make singular decomposition on $X_1^T Y_1$:

$$X_1^T Y_1 = U[\textstyle\sum 0]V^T \quad (6)$$

where $\sum = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)$, $U \in R^{p \times p}$ and $V \in R^{q \times q}$ are orthogonal matrices.

$$X_1 U = [\boldsymbol{u}_1', \boldsymbol{u}_2', \cdots, \boldsymbol{u}_p'],$$
$$Y_1 V = [v_1', v_2', \cdots, v_q'] \quad (7)$$

According to Eq. (6), we can define the angle $\theta_1, \theta_2, \cdots, \theta_p \in [0, \frac{\pi}{2}]$ between the S_1 and S_2:

$$\cos\theta_k = \boldsymbol{u}_k'^T v_k' = \sigma_k \quad k = 1,2,\cdots,p \quad (8)$$

obviously, we have $0 \le \theta_1 \le \cdots \le \theta_p \le \frac{\pi}{2}$. We call $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_p\}$ and $\{v_1, v_2, \cdots, v_p\}$ as principal eigen vectors of the subspace pair (S_1, S_2).

According to Eq. (8), we may define the distance measure between the S_1 and S_2:

$$\mathrm{dist}(S_1,S_2) = \sqrt{\sum_{i=1}^{p}(1-\cos^2\theta_i)} = \sqrt{\sum_{i=1}^{p}(1-T_i^2)} \quad (9)$$

According to above definition, we attain following conclusions for the subspace distance measure:

(1) Nonnegativity: $\mathrm{dist}(S_1,S_2) \geq 0$

(2) Symmetrization: $\mathrm{dist}(S_1,S_2) = \mathrm{dist}(S_2,S_1)$

Definition 1: For an arbitrary matrix $A = [a_{ij}] \in R^{m\times d}$, we define the F-norms (Frobeuius norms) of the matrix A:

$$||A||_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{d}|a_{ij}|^2\right)^{\frac{1}{2}} = (tr(A^TA))^{\frac{1}{2}} \quad (10)$$

where $tr(.)$ stands for the matrix trace. Generally, F-norms possesses orthogonal invariance, i.e., if Q and Z are orthogonal matrices, then:

$$||QAZ||_F = ||A||_F \quad (11)$$

Theorem 1: Assume that $X_1 \in R^{d\times p}$ and $Y_1 \in R^{d\times q}$ are the column matrices composed of orthogonal basis vectors of the subspace S_1 and S_2, using Eqs. (6), (7) and (8), we obtain:

$$\mathrm{dist}(S_1,S_2) = \sqrt{tr(X_1^TY_2Y_2^TX_1)} = ||X_1^TY_2||_F \quad (12)$$

where, $X = [X_1, X_2]$ and $Y = [Y_1, Y_2]$ are orthogonal matrix in $R^{d\times d}$. The proof see Appendix A.

Theorem 1 is called distance measure theorem, which provides us several methods computing the distance between two subspaces.

Corollary: The distance between two subspaces S_1, S_2 is bounded, i.e., $0 \leq \mathrm{dist}(S_1,S_2) \leq \sqrt{p}$, if and only if S_1 and S_2 coincide, becomes 0; if and only if S_1 and S_2 are orthogonal, becomes $\sqrt{p}$.

3.2 The Orthogonal Iterating Convergence Theorem for the LSMs[9)]

Given that symmetric matrix $A \in R^{d\times d}$, suppose that $Q_0 \in R^{d\times p}$ is an orthogonal column matrix composed of given initial basis vectors, then use the orthogonal iterating method to produce a matrix sequence $\{Q_k\} \subset R^{d\times p}$ as follows:

$$\begin{cases} Z_k = AQ_{k-1} \\ Q_kR_k = Z_k \end{cases} \quad k = 1,2,\cdots \quad (13)$$

where Z_k is a transformed matrix of the orthogonal matrix Q_{k-1}; R_k is an upper triangular matrix which includes positive element, resulting in the Z_k being transformed by Gram-Schmidt orthogonality normalization. In order to analyse the nature of this iterating method, assume:

$$Q^TAQ = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_d) \triangleq D \qquad |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_d| \quad (14)$$

where

$$Q = [Q_\alpha, Q_\beta] \qquad D = \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} \quad (15)$$

and $D_1 \in R^{p\times p}$, $D_2 \in R^{(d-p)\times(d-p)}$ are diagonal matrix, $Q_\alpha \in R^{d\times p}$, $Q_\beta \in R^{d\times(d-p)}$ are the block matrices of the Q, λ_i is the ith eigen value. If $|\lambda_p| > |\lambda_{p+1}|$, we call the $R(Q_\alpha)$ as p-dimensional principal eigen subspace of the matrix A, which is also invariant one corresponding to the principal eigen value $\lambda_1, \lambda_2, \cdots, \lambda_p$.

Suppose that the orthogonal iterating method is convergent, then:

$$\lim_{k\to\infty} R(Q_k) = R(Q_\alpha) \quad (16)$$

Below, we give a theorem for the p-dimensional principal eigen subspace of the matrix A to which the subspace $R(Q_k)$ produced by the orthogonal iterating method converges.

Theorem 2: Assume that the eigenvalue decomposition of the symmetric matrix $A \in R^{d\times d}$ is given by Eqs. (14) and (15), and $|\lambda_p| > |\lambda_{p+1}|$, $s = \mathrm{dist}[R(Q_\alpha), R(Q_0)] < 1$, then the matrix sequence Q_k formed by Eq. (13) meets:

$$\mathrm{dist}[R(Q_\alpha), R(Q_k)] \leq \sqrt{\frac{\frac{(d-p)s^2}{p-s^2}|\frac{\lambda_{p+1}}{\lambda_p}|^{2k}}{1+\frac{(d-p)s^2}{p(p-s^2)}|\frac{\lambda_{p+1}}{\lambda_p}|^{2k}}} \quad (17)$$

thus

$$\lim_{k\to\infty} R(Q_k) = R(Q_\alpha) \quad (18)$$

The proof see Appendix B.

According to Eq. (17), the convergence speed which the subspace $R(Q_k)$ converges to $R(Q_\alpha)$ is related to $|\frac{\lambda_{p+1}}{\lambda_p}|$, the bigger the value, the faster the convergence speed. This theorem indicates that the above orthogonal iterating method can form the eigen subspace of the symmetric matrix A.

4. The Robustness and Disturbance for the LSMs

Assume that the general form for the self-supervised LSMs is as follows[9),11)]:

$$L_k = (I + \mu_k \boldsymbol{x}_k \boldsymbol{x}_k^T)L_{k-1} = A_k L_{k-1}$$
$$= A_k A_{k-1} \cdots A_1 L_0 = \widehat{A}_k L_0 \quad (19)$$

where $A_k \triangleq I + \mu_k \boldsymbol{x}_k \boldsymbol{x}_k^T$ is called a transformed matrix at the kth instant; $\widehat{A}_k = A_k A_{k-1} \cdots A_1$ a transformed matrix form the $k = 1$ instant to the kth instant.

Theorem 3: For the LSM of Eq. (19), assume $\mu_k = \mu \ll \frac{2}{N-1}$ and $||\boldsymbol{x}_k|| = 1$ (without loss of generality), then the convergence solution for $\widehat{A}_k$ is:

$$A = \lim_{k \to N} \widehat{A}_k \approx I + \mu' \widehat{R} \quad (20)$$

where $\widehat{R} = \frac{1}{N} \sum_{m=1}^{N} \boldsymbol{x}_m \boldsymbol{x}_m^T$ is the estimate of pattern autocorrelation matrix; $\mu' = N\mu$.

Proof: Let $Z_k = \boldsymbol{x}_k \boldsymbol{x}_k^T$

$$\because \widehat{A}_k = (I + \mu Z_k)(I + \mu Z_{k-1}) \cdots (I + \mu Z_1)$$
$$= I + \mu \sum_{m=1}^{k} Z_m + \mu^2 \sum_{m_1=2}^{k} \sum_{m_2=1}^{m_1-1} Z_{m_1} Z_{m_2}$$
$$+ \mu^3 \sum_{m_1=3}^{k} \sum_{m_2=2}^{m_1-1} \sum_{m_3=1}^{m_2-1} Z_{m_1} Z_{m_2} Z_{m_3} + \cdots$$
$$+ \mu^k \sum_{m_1=k}^{k} \sum_{m_2=k-1}^{m_1=1} \cdots \sum_{m_k=1}^{m_k-1} Z_{m_1} Z_{m_2} \cdots Z_{m_k} \quad (21)$$

Because of $||\boldsymbol{x}_k|| = 1$, the trace of the Z_{m_k} is:

$$tr(Z_{m_k}) = tr(\boldsymbol{x}_{m_k} \boldsymbol{x}_{m_k}^T)$$
$$= tr(\boldsymbol{x}_{m_k}^T \boldsymbol{x}_{m_k}) = 1 \quad (22)$$
$$\therefore tr(Z_{m_1} Z_{m_2} \cdots Z_{m_k}) = 1 \quad (23)$$

In order to derive the approximate solution of the transformed matrix, general method is to discuss the approximate solution corresponding to its trace, so we make the trace the $\widehat{A}_k$:

$$tr(\widehat{A}_k) = d + \mu k + \mu^2 C_k^2 + \cdots + \mu^k C_k^k$$
$$= d + \mu k \left(1 + \mu \frac{k-1}{2} + \mu^2 \frac{(k-1)(k-2)}{6} + \cdots + \frac{\mu^{k-1}}{k}\right) \quad (24)$$

Therefore, if $\mu \frac{N-1}{2} \ll 1$, i.e., $\mu \ll \frac{2}{N-1}$, when $k \to N$, we can neglect the term of above μ^2.

$$\therefore A = \lim_{k \to N} \widehat{A}_k \approx I + \mu \sum_{m=1}^{N} Z_m$$
$$= I + \mu N \frac{1}{N} \sum_{m=1}^{N} \boldsymbol{x}_m \boldsymbol{x}_m^T \triangleq I + \mu' \widehat{R} \quad (25)$$

where $\mu' = N\mu$, $\widehat{R} = \frac{1}{N} \sum_{m=1}^{N} \boldsymbol{x}_m \boldsymbol{x}_m^T$. □

Obviously, when the N pattern samples are orthogonal, the high order term (above μ^2) in the right of Eq. (24) are all zero, so Eq. (20) becomes an accurate expression .

The above result can be generalized to any LSMs, i.e., $\widehat{R}$ can be varied with different learning methods[9)]:

$$\widehat{R} = \mu^{(i,i)} \widehat{R}^{(i)} - \sum_{j \neq i} \mu^{(i,j)} \widehat{R}^{(j)} \quad (26)$$

where $\widehat{R}^{(i)}$ is the auto-correlation matrix for the ith category subspace, $\mu^{(i,j)}$ is the learning coefficient for the patterns in the jth category rotating or learning the ith category subspace.

Lemma 1: Given that a symmetric matrix $A = I + \tau \boldsymbol{c}\boldsymbol{c}^T \in R^{d \times d}$, where the 2-norm of the $\boldsymbol{c} \in R^d$ is 1, and $\tau \in R$. Then[9)]:

(1) $\lambda_i(A) = 1 + \alpha_i \tau$, $\alpha_i > 0$, $\sum_{i=1}^{d} \alpha_i = 1$.
(2) If $\tau > 0$, $\lambda_i(A) \geq 1$, $i = 1, 2 \cdots, d$.
(3) If $\tau \leq 0$, $\lambda_i(A) \leq 1$, $i = 1, 2, \cdots, d$.

where $\lambda_i(A)$ is the ith eigen spectrum of the matrix A.

According to this lemma and the properties of the eigen spectrum of a matrix, we easily obtain the eigen spectrum distribution theorem corresponding to the transformed matrix of the self-supervised LSM:

Theorem 4: The eigen spectrum corresponding to the transformed matrix of a subspace $A^{(i)} = I + \mu^{(i,i)} \widehat{R}^{(i)} - \sum_{j \neq i} \mu^{(i,j)} \widehat{R}^{(j)}$ is[9),11)]:

$$\lambda_k(A^{(i)}) = 1 + \mu^{(i,i)} \sum_{m=1}^{N_i} \alpha_m^{(i)}(k)$$
$$- \sum_{j \neq i} \mu^{(i,j)} \sum_{m=1}^{N_j} \alpha_m^{(j)}(k) \quad (27)$$

where

$$\sum_{k=1}^{d} \alpha_m^{(i)}(k) = 1, \quad \sum_{k=1}^{d} \alpha_m^{(j)}(k) = 1$$

This theorem is easily induced from the above Lemma 1.

From above Theorem 4, we can see that the mechanism of the self-supervised LSMs is to gradually increase the eigen values of principal components in its own subspace, and to progressively reduce the effects of the eigen spectrums from the other category subspaces. Below we discuss the disturbed properties of the self-supervised LSMs.

Theorem 5: Assume that the transformed matrix $A \in R^{d\times d}$ is disturbed by a random matrix $E \in R^{d\times d}$, the decomposing and the blocking of the matrix A and E are as follows[9),11)]:

$$A = Q \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix} Q^T$$

$$Q^T E Q = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \quad (28)$$

where $Q = [Q_1, Q_2]$ is an orthogonal matrix, $Q_1 \in R^{d\times p}$, $Q_2 \in R^{d\times(d-p)}$, and assume $2p \le d$, and suppose that there exists a matrix $P \in R^{(d-p)\times p}$, which ensures that the column vectors of $\widehat{Q}_1 = (Q_1 + Q_2 P)(I + P^T P)^{-\frac{1}{2}}$ are the orthogonal basis ones of $R(A+E)$, then the distance for their eigen subspaces:

$$\mathrm{dist}[R(Q_1), R(\widehat{Q}_1)] \le \frac{2||E_{21}||_F}{\sigma} \quad (29)$$

where $\sigma = \lambda_p - \lambda_{p+1} - ||E_{11}||_F - ||E_{22}||_F > 0$ is called separation degree, $||E_{12}||_F \cdot ||E_{21}||_F \le \frac{\sigma^2}{4}$.

Proof: From the above given condition, Ref. 12) has proved: there exists a matrix $P \in R^{(d-p)\times p}$:

$$||P||_F \le \frac{2||E_{21}||_F}{\sigma} \quad (30)$$

so that each column vector in matrix $\widehat{Q}_1 = (Q_1 + Q_2 P)(I + P^T P)^{-\frac{1}{2}}$ forms a set of orthogonal base ones corresponding to eigen subspace of matrix $A + E$. So we obtain:

$$Q_1^T \widehat{Q}_1 = (I + P^T P)^{-\frac{1}{2}} \quad (31)$$

thus

$$||Q_1^T \widehat{Q}_1||_F^2 = \sum_{i=1}^{F} \frac{1}{1 + \sigma_i^2(P)} \quad (32)$$

where $\sigma_i(P) > 0$ indicates the ith eigen value for matrix P which is usually very small. Generally, $\max\{\sigma_1(P)\} \ll 1$, then:

$$\frac{1}{1 + \sigma_i^2(P)} \approx 1 - \sigma_i^2(P) \quad (33)$$

$$\therefore ||Q_1^T \widehat{Q}_1||_F^2 \approx \sum_{i=1}^{F} (1 - \sigma_i^2(P)) = p - ||P||_F^2 \quad (34)$$

Finally, according to Theorem 1, we obtain:

$$\mathrm{dist}[R(Q_1), \widehat{R}(Q_1)] = \sqrt{p - ||Q_1^T \widehat{Q}_1||_F^2} \approx ||P||_F \le \frac{2||E_{21}||_F}{\sigma} \quad (35)$$

□

This theorem indicates that the upper bound of the distance between the disturbed eigen subspace and the old eigen subspace is associated with the disturbed properties in random matrix E and the difference between the smallest eigen value λ_p in its own eigen subspace and the largest eigen value λ_{p+1} in the corresponding complementary subspace; which discloses that the learning mechanism of the self-supervised LSMs successively increases the principal eigen values in pattern eigen subspace by means of repeatedly iterating, so that the value $\lambda_p - \lambda_{p+1}$ becomes bigger and bigger. Finally, the disturbed upper bound of the pattern eigen subspace formed is inversely proportional to the separation degree between the eigen spectrum corresponding to the principal eigen subspace and the eigen spectrum corresponding to its complementary eigen subspace[9),11)].

5. Conclusions

This paper discusses the statistical and disturbed properties for self-supervised learning subspace methods for pattern recognition. First, the new distance measure between two subspaces is defined, the distance measure's theorem based on the F-norms is given and proved. We prove the convergence of the learning subspace methods by two steps. The first step is to prove that the iterating subspace is convergent, the second step that the convergence solution of the iterating subspace is also convergent by an orthogonal iterating method. So the orthogonal iterating convergent theorem for subspaces is given and proved. Afterwards, the robustness and disturbance properties for the self-supervised LSMs are discussed, the approximate expression of the eigen spectrum caused by LSM's rotation is also given. Finally, the upper bound disturbed theorem for the LSMs is gived and proved.

From the analyses of the above statistical and disturbed properties, we can attain several conclusions about generalization capabilities as follows:

(1) For self-supervised clustering methods, their generalization capabilities are associated with the learning algorithm used, and the qual-

ity and the quantity of training samples available which make a main effect on the robustness and disturbance of the clustering subspace formed.

(2) The working mechanism of the self-supervised LSMs is to gradually increase the eigen values of principal components in its own subspace, and to progressively reduce the effects of the eigen spectrums from the other categories' subspaces.

(3) The upper bound of the distance between the disturbed eigen subspace and the old subspace is associated with the disturbed properties in random matrix E, and the difference between the smallest eigen value λ_p in its own eigen subspace and the largest eigen value λ_{p+1} in corresponding complementary subspace.

Obviously, above analysis of the disturbed properties and generalization capabilities for self-supervised learning subspace methods for pattern recognition is instructive to speech recognition and clutter classification in time varying and nonstationary dynamic environments.

References

1) Watanabe, S., Lambert, P.F., Kulikowski, C.A., et al.: *Computer and Information Sciences* Vol.2, Academic Press, New York, pp.91–122 (1967).

2) Kittler, J., Trappl, R., Klir, G.J. and Ricciardi, L. (Eds.): *Progress in Cybernetics and Systems Research*, Vol.3, Hemisphere Publ., Washington DC, pp.92–97 (1978).

3) Noguchi, Y.: *Bull. of the Electrotech. Lab.* Vol.40, No.7, pp.11–27 (1976).

4) Kohonen, T.: *Self-organization and Associative Memory*, Springer-Verlag, Berlin Heidelberg New York, 1989.

5) Oja, E. and Kuuusela, M.: The ALSM Algorithm – An Improved Subspace Method of Classification, *Pattern Recognition*, Vol.16, No.4, pp.421–427 (1983).

6) Oja, E.: *Subspace Methods of Pattern Recognition*, Research Studies Press (1983).

7) Huang D.S. and Bao Z.: The Improved Learning Subspace Method for the Recognition of Objects – Ships, Chaff, *IJCNN'92*, Beijing (Oct. 1992).

8) Huang, D.S. and Bao Z.: The Forward and Backward Smoothing Learning Subspace Method for Pattern Recognition, *IC of AMSE on MSC'92*, Hefei (Nov. 1992).

9) Huang D.S.: The Study of Recognition Technique Based on One-dimensional Image of Radar Targets, Ph.D. Dissertation, Xidian University, China (Dec. 1991).

10) Huang D.S.: A Study on Three Kinds of New Self-Supervised Learning Subspace Methods for Pattern Recognition, *Acta Elctronica Sinica*, Vol.23, No.7, pp.25–30 (1995).

11) Huang D.S.: An Analysis of the Statistical Properties on the Self-Supervised Learning Subspaces for Pattern Recognition, *Acta Elctronica Sinica*, Vol.23, No.9, pp.99–102 (1995).

12) Golub, G.H. and Van Loan, C.F.: *Matrix Computations*, The Johns Hopkins University Press (1988).

Appendix A

Proof:

$$\because [Y_1 \quad Y_2]\begin{bmatrix} Y_1^T \\ Y_2^T \end{bmatrix} = I \qquad \text{(A-1)}$$

$$\therefore Y_1 Y_1^T = I - Y_2 Y_2^T \qquad \text{(A-2)}$$

$$I - X_1^T Y_1 Y_1^T X_1 = X_1^T Y_2 Y_2^T X_1 \qquad \text{(A-3)}$$

$$tr(X_1^T Y_2 Y_2^T X_1) = tr(I - X_1^T Y_1 Y_1^T X_1) \qquad \text{(A-4)}$$

From Eq. (6), we obtain:

$$X_1^T Y_1 Y_1^T X_1 = U \sum{}^2 U^T \qquad \text{(A-5)}$$

$$\begin{aligned} \therefore\ & tr(I - X_1^T Y_1 Y_1^T X_1) \\ &= tr(U(I - \textstyle\sum^2)U^T) \\ &= tr(diag(1-\sigma_1^2, 1-\sigma_2^2, \cdots, 1-\sigma_p^2)) \\ &= \sum_{i=1}^{p}(1-\sigma_i^2) \end{aligned} \qquad \text{(A-6)}$$

$$\therefore \mathrm{dist}(S_1, S_2) = \sqrt{(tr(X_1^T Y_2 Y_2^T X_1)} \qquad \text{(A-7)}$$

Besides

$$||X_1^T Y||_F^2 = ||X_1^T Y_1||_F^2 + ||X_1^T Y_2||_F^2 \qquad \text{(A-8)}$$

and

$$\begin{aligned} ||X_1^T Y||_F^2 &= ||X_1||_F^2 = tr(X_1^T X_1) \\ &= tr(I_p) = p \end{aligned} \qquad \text{(A-9)}$$

$$||X_1^T Y_1||_F^2 = \sum_{i=1}^{p}\sigma_i^2 \qquad \text{(A-10)}$$

Combining Eq. (A-8) with Eq. (A-9):

$$||X_1^T Y_2||_F = \sqrt{||X_1^T Y||_F^2 - ||X_1^T Y_1||_F^2} = \sqrt{p - \sum_{i=1}^{p} \sigma_i^2} \quad \text{(A-11)}$$

$$\therefore \text{dist}(S_1, S_2) = ||X_1^T Y_2||_F \quad \text{(A-12)}$$

This completes the proof for Theorem 1.

Appendix B

Proof: From Eq. (13), we deduce:

$$A^k Q_0 = Q_k (R_k R_{k-1} \cdots R_1) \quad \text{(B-1)}$$

From Eqs.(14) and (15), we obtain:

$$Q^T A^k Q = D^k = \begin{bmatrix} D_1^k & 0 \\ 0 & D_2^k \end{bmatrix} \quad \text{(B-2)}$$

$$\therefore D^k Q^T Q_0 = Q^T Q_k (R_k R_{k-1} \cdots R_1) \quad \text{(B-3)}$$

$$D^k \begin{bmatrix} V_0 \\ W_0 \end{bmatrix} = \begin{bmatrix} V_k \\ W_k \end{bmatrix} (R_k R_{k-1} \cdots R_1) \quad \text{(B-4)}$$

Let $V_k \triangleq Q_\alpha^T Q_k$, $W_k \triangleq Q_\beta^T Q_k$, $k = 0, 1, 2, \cdots$. Combining Eqs. (B-4) and (B-2):

$$\begin{bmatrix} D_1^k & 0 \\ 0 & D_2^k \end{bmatrix} \begin{bmatrix} V_0 \\ W_0 \end{bmatrix} = \begin{bmatrix} V_k \\ W_k \end{bmatrix} (R_k R_{k-1} \cdots R_1) \quad \text{(B-5)}$$

Because general V_0 is nonsingular, the solution of this equation:

$$W_k = D_2^k W_0 V_o^{-1} D_1^{-k} V_k \quad \text{(B-6)}$$

Using the distance measure given by Theorem 1 ($p = q$)

$$\therefore \text{dist}[R(Q_\alpha), R(Q_k)] = ||Q_\beta^T Q_k||_F = ||W_k||_F \le \frac{||D_2^k||_F}{||D_1^k||_F} \frac{||W_0||_F}{||V_0||_F} ||V_k||_F \quad \text{(B-7)}$$

$$\because ||D_2^k||_F = \sqrt{\sum_{i=p+1}^{d} |\lambda_i|^{2k}} \le \sqrt{d-p}\,|\lambda_{p+1}|^k \quad \text{(B-8)}$$

$$||D_1^k||_F = \sqrt{\sum_{i=1}^{p} |\lambda_i|^{2k}} \ge \sqrt{p}\,|\lambda_p|^k \quad \text{(B-9)}$$

$$\because \text{dist}[R(Q_\alpha), R(Q_0)] = ||Q_\beta^T Q_0||_F = s$$

$$\therefore ||Q_\alpha^T Q_0||_F = \sqrt{p - s^2} \quad \text{(B-10)}$$

In the light of the same method:

$$||V_k||_F = ||Q_\alpha^T Q_k||_F = \sqrt{p - ||Q_\beta^T Q_k||_F^2} = \sqrt{p - ||W_k||_k^2} \quad \text{(B-11)}$$

Combining Eqs. (B-8), (B-9), (B-10) and (B-11) with (B-7), we obtain:

$$||V_k||_F \le \sqrt{\frac{d-p}{p}} \left|\frac{\lambda_{p+1}}{\lambda_p}\right|^k \times \frac{s}{\sqrt{p - s^2}} \sqrt{p - ||W_k||_F^2} \quad \text{(B-12)}$$

The solution of this inequality is Eq. (17).

Because of $|\lambda_p| > |\lambda_{p+1}|$, when $k \to \infty$, $\text{dist}[R(Q_\alpha), R(Q_k)] \to 0$.

That is:

$$\lim_{k \to \infty} R(Q_k) = R(Q_\alpha) \quad \text{(B-13)}$$

This proof is complete.

(Received December 26, 1994)
(Accepted January 10, 1996)

D.S. Huang was born in Anhui Province, P.R. China, January 11, 1964. He received the M.S. and Ph.D. degrees in electrical engineering from the National University of Defense Technology, Changsha, P.R. China and Xidian University, Xian, P.R. China, in 1989 and 1993, respectively. He became a post doctoral research fellow at the Beijing Institute of Technology, Beijing, P.R. China, from July 1993 to July 1995. He is now a post doctoral research fellow at the National Laboratory of Pattern recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, P.R. China. Research interests include signal processing, pattern recognition, neural networks and computer vision.