Pattern Recognition Letters 4 (1986) 63-69
North-Holland

April 1986

A generalised learning algorithm based on guard zones

A. PATHAK and S.K. PAL

Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta 700 035, India

Received 16 August 1985
Revised 8 January 1986

Abstract: An algorithm for learning class parameters using a restricted updating programme is described along with investigation of its convergence for optimum learning. The algorithm is a generalisation of some existing ones which were found to be useful for practical data.

Key words: Guard zone, stochastic convergence, learning, robust estimates.

1. Introduction

An adaptive pattern recognition system can be viewed as a learning machine in which the decision of the system gradually approaches the optimal decision by acquiring necessary information from observed patterns. System performance is improved as a result. In a supervised system, the machine requires an extra source of knowledge, usually of a higher order, for correcting the decision taken by a classifier. When an extra source of knowledge on which a supervisory programme could be based is not readily available, the performance of the system becomes highly unpredictable.

The most widely used tools for recursive learning of class parameters are Bayesian estimation methods and stochastic approximation (Fu, 1968; Tsypkin, 1973). In this context, we would like to single out the self-supervised learning system based on the concept of a 'guard zone', mooted by Pal et al. (1980). It was used to restrict the updating of estimates of parameters (feature means and variances) by means of 'doubtful' samples. For this purpose a guard zone was defined for each class in such a way that a training sample was used for updating only if it fell within the guard zone. A similar algorithm was presented by Chien (1970) as a solution to the problem of identifying 'spurious', that is, possibly non-representative training samples for the case when feature means are to be learned. A threshold is defined such that if the 'distance' of the current training sample from the preceding estimate of the mean (the same 'distance' is used for defining a guard zone) exceeds it, the training sample is rejected. As such both algorithms are basically the same, and are stochastic approximation procedures of sorts.

Although these algorithms were tested with success on some practical data, the two works did not provide any proof of convergence or any theoretical investigation of the choice of the controlling parameter namely, guard zone dimension or threshold for optimum learning.

The present work describes a generalized version of the two, called the Generalized Guard Zones Algorithm (GGA). Basically, it aims to detect outliers and reject them from the parameter-updating procedure. As such, it can be looked upon as a robust estimation procedure (Andrews, 1975; Huber, 1981). (Essentially, the term 'robustness' signifies insensitivity to small de' tions from the underlying assumptions (

1981)). Under rather general assumptions, we have investigated the stochastic convergence of this algorithm for some special problems of estimation. For this purpose, results on stochastic approximation are used.

2. The Generalized Guard Zones Algorithm (GGA)

Let $X=[X_1, X_2, \ldots, X_N]'$, $X \in \mathbb{R}^N$ be an N-dimensional feature vector defined over a pattern class C.

Let us make the following assumptions:

(A1) The distribution of X over C is continuous.
(A2) This distribution depends on a q-dimensional parameter vector θ, some or all of which need to be learned.
(A3) The distribution of X over C is such that $E(X)$ exists and is equal to μ.
(A4) The dispersion matrix of X, namely,

$$\mathrm{Disp}(X)=\Sigma=((\sigma_{ij})) \text{ exists.}$$

Before stating the algorithm itself, let us define a guard zone formally as follows:

Definition. Let S be a metric space and δ a metric defined on it. Then for any point $a \in S$, a guard zone $G(a, \lambda)$ having an 'extent' λ is the subset of S defined by

$$G(a, \lambda)=\{x : \delta(a, x) \le \lambda\},$$

where $\lambda \ge 0$.

Clearly, $G(a, \lambda)$ is nothing but a closed ball of radius λ centred at a in S with respect to the metric δ.

In the subsequent discussions we shall be taking $S=\mathbb{R}^N$ and a metric d defined as

$$d^2(x, y)=(x-y)'A(x-y), \quad x, y \in \mathbb{R}^N,$$

A being a symmetric, positive definite matrix.

Let us now proceed to the algorithm properly.

Let $X_1, X_2, X_3, \ldots$ be the sequence of learning (or training) samples, randomly selected from C, that is, assumed to be independently and identically distributed, i.e., we assume that correctly labelled training samples are available.

We restrict ourselves to the case where θ includes μ and/or elements of Σ only.

The generalized guard zones algorithm (GGA) for estimating θ recursively is as follows:

$$\theta_k=\begin{cases} f(X_k) & \text{for } k=1, \\ \theta_{k-1}-a_k Y_k & \text{for } k>1; \end{cases} \tag{1}$$

$$Y_k=\begin{cases} \theta_{k-1}-f(X_k) & \text{if } X_k \in G(m_{k-1}, \lambda_k), \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

θ_k: the k-th-stage estimate of θ.

$\{a_k\}$: a sequence of positive numbers, with $a_k \le 1 \ \forall k$.

f: $\mathbb{R}^N \to \mathbb{R}^q$ is a continuous mapping, defining an unbiased statistic for θ.

m_{k-1}: the $(k-1)$-th stage GGA estimate of μ.

$$G(m_{k-1}, \lambda_k)=\{x : x \in \mathbb{R}^N, d_k(x, m_{k-1}) \le \lambda_k\},$$

$$d_k^2(x, y)=(x-y)'A_k(x-y),$$

A_k: A symmetric, positive definite matrix, which may or may not be a function of X_i and/or θ_i, $i=1(1)k$.

λ_k: A positive number, prespecified.

In essence, this algorithm uses only those training samples for updating the estimate, which lie within the corresponding guard zone centred at the preceding estimate of the mean. Training samples which lie outside it are ignored and the estimate kept unchanged at the corresponding stages.

Special cases

(1) When $q=N$, $\theta=\mu$, i.e., only the mean vector is to be learned, we have $f(X)=X$, and the algorithm is as follows:

$$\theta_1=m_1=X_1,$$

and for $k>1$, with $\theta_k=m_k$,

$$m_k=\begin{cases} m_{k-1}-a_k[m_{k-1}-X_k] \\ \quad \text{if } d_k(m_{k-1}, X_k) \le \lambda_k, \\ \mu_{k-1} \quad \text{otherwise.} \end{cases} \tag{3}$$

(2) For estimating σ_{ij}, $i=1(1)N$, $j=1(1)N$, recursively, there are two alternative procedures (we write $\hat{\sigma}_{ij(k)}=s_{ij}^{(k)}$):

(a) $s_{ij}^{(1)}=0$, and for $k>1$,

$$s_{ij}^{(k)} = \begin{cases} s_{ij}^{(k-1)} - a_k[s_{ij}^{(k-1)} - (X_{k_i} - m_{(k-1)_i})(X_{k_j} - m_{(k-1)_j})] \\ \quad \text{if } d_k(m_{n-1}, X_k) \le \lambda_k, \\ s_{ij}^{(k-1)} \quad \text{otherwise.} \end{cases} \tag{4a}$$

(b) $s_{ij}^{(1)} = 0$, and for $k > 1$,

$$s_{ij}^{(k)} = c_{ij}^{(k)} - m_{(k-1)_i} m_{(k-1)_j}, \tag{4b}$$

where

$$c_{ij}^{(k)} = \begin{cases} c_{ij}^{(k-1)} - a_k[c_{ij}^{(k-1)} - X_{k_i} X_{k_j}] \\ \quad \text{if } d_k(m_{n-1}, X_k) \le \lambda_k, \\ c_{ij}^{(k-1)} \quad \text{otherwise.} \end{cases} \tag{5}$$

(3) When $\theta = [\mu' \vdots \sigma']'$, $\sigma' = [\sigma_{11}\, \sigma_{22} \cdots \sigma_{NN}]$, $q = 2N$,

$$A_k = \begin{pmatrix} s_{11} & 0 & 0 & \cdots & 0 \\ 0 & s_{22} & 0 & \cdots & 0 \\ 0 & 0 & s_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & s_{NN} \end{pmatrix}^{-1}$$

$$= [\text{Diag}(s_{11}, s_{22}, \ldots, s_{NN})]^{-1},$$

$$a_k = \frac{1}{k}, \qquad \lambda_k = \frac{1}{\lambda},$$

then the GGA reduces to the algorithm of Pal et al. (1980).

(4) When $\theta = \mu$, $q = N$,

$A_k = K_1$, the dispersion in X due to ordinary measurement variation, i.e., type 1 noise (see Chien (1970)),

$\lambda_k = \theta_k$ (the threshold in Chien's algorithm),

$a_k = [(k-1) + \nu]^{-1}$ where ν is such that

$$\text{Disp}(\mu_0) = K_1/\nu,$$

$\mu_0 =$ being the initial estimate, $\nu > 0$,

the GGA reduces to the non-linear learning algorithm of Chien (1970).

(5) As all λ_k's decrease progressively, the system approaches the nonadaptive state. Clearly, this is because the dimensions of the guard zones decrease and hence the probability of a training sample to fall within the guard zone decreases with decrease in the corresponding λ_k-value; so the number of training samples getting selected for the updating process decreases.

(6) On the other hand, when the λ_k's increase progressively, the system approaches the non-supervised state for, as the 'extent' of the guard zones increases, more and more training samples get selected for the updating of estimates; that is, the updating programme becomes less and less restrictive.

3. Convergence of the generalized guard zones algorithm

The convergence of a recursive discrete algorithm for estimating a parameter θ by $\hat{\theta}_n$, can be defined in various ways. For instance, we say that

(i) the sequence $\{\hat{\theta}_n\}$ converges to θ with probability one or almost surely if

$$P\left[\lim_{n \to \infty} |\hat{\theta}_n - \theta| = 0\right] = 1,$$

P being the probability measure.

(ii) $\{\hat{\theta}_n\}$ converges to θ in the mean-square sense if

$$\lim_{n \to \infty} E[|\hat{\theta}_n - \theta|^2] = 0,$$

E being the expectation operator.

For proving certain results on the convergence of the GGA, we shall be making use of the following results:

Theorem 1 (Schmetterer, 1968). *Let $\{a_n\}$ be a sequence of positive real numbers such that*

$$\text{(B1)} \quad \sum_{n=1}^{\infty} a_n^2 < \infty.$$

Let x_n and y_n be K-dimensional random vectors which satisfy

$$\text{(B2)} \quad x_{n+1} = x_n - a_n y_n, \quad n \ge 1.$$

Let M_n be a measurable mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$, such that

$$\text{(B3)} \quad E(y_n | x_1, x_2, \ldots, x_n) = M_n(x_n) \text{ a.e.}$$

Let a, b, c be nonnegative real numbers and let

$$\text{(B4)} \quad E(|y_n|^2 | x_1, x_2, \ldots, x_n) \le a + b|x_n| + c|x_n|^2 \text{ a.e.}$$

Also, for every $x \in \mathbb{R}^K$ and $n \ge 1$,

(B5) $x'M_n(x) \geq 0$.

If x_1 is chosen in such a way that

(B6) $E(|x_1|^2)$ *exists,*

then the sequence $\{x_n\}$ converges with probability 1 (that is, almost surely) and the sequence $\{E|x_n|^2\}$ converges also.

Theorem 2 (Schmetterer, 1968). *Suppose that assumptions* (B1)–(B6) *hold. If there exists for every $\eta > 0$ a $\delta > 0$ such that for $n \geq 1$*

$$\text{(B7)} \quad \inf_{\eta < |x| < \eta'} |x'M_n(x)| \geq \delta,$$

then $\{x_n\}$ converges to the k-dimensional null vector $\mathbf{0}$ almost surely.

We shall now be proving the following:

Proposition 1. *For the problem of estimating $\theta = \mu$ recursively, let $\hat{\theta}_k = m_k$ be the sequence of estimates, where m_k is given by equation* (3).

If

$$\text{(C1)} \quad \sum_{n=1}^{\infty} a_n^2 < \infty$$

and

$$\text{(C2)} \quad p_k = P[d_k(m_{k-1}, X_k) \leq \lambda_k \mid m_{k-1}] > \delta, \quad \forall k,$$

for some $\delta > 0$,

then

(a) $\{m_k\}$ *converges with probability 1 to μ as $k \to \infty$;*

(b) $\{E|m_k - \mu|^2\}$ *converges as $k \to \infty$.*

Proposition 2. *Consider the problem of estimating μ and Σ completely.*

Let

$$\underset{q \times 1}{\theta'} = [\mu' \vdots \sigma^{*(1)'} \vdots \sigma^{*(2)'} \vdots \cdots \vdots \sigma^{*(N)'}] \tag{6a}$$

and

$$\underset{q \times 1}{\hat{\theta}_k'} = [m_k' \vdots c_k^{(1)'} \vdots c_k^{(2)'} \vdots \cdots \vdots c_k^{(N)'}], \tag{6b}$$

where

$$q = \frac{N(N+1)}{2},$$

$$\underset{q_i \times 1}{\sigma^{*(i)'}} = [\sigma_{ii}^* \; \sigma_{i(i+1)}^* \cdots \sigma_{iN}^*],$$

$$q_i = N - i + 1,$$

$$\sigma_{ij}^* = E(X_i X_j) = \sigma_{ij} + \mu_i \mu_j,$$

and the elements of m_k and $c_k^{(i)} = [c_{ii}^{(k)} c_{i(i+1)}^{(k)} \cdots c_{iN}^{(k)}]$ are given by equations (3) *and* (5) *respectively.*

If the conditions (C1) *and* (C2) *hold, and besides, the following condition is also satisfied:*

(C3) $\eta_{ij} = E(X_i^2 X_j^2)$ *exists* $\forall\ i, j = 1(1)N$ *for the elements of the feature vector X,*

then

(a) $\{\hat{\theta}_k\}$ *(given by* (6b)*) converges with probability 1 to θ (given by* (6a)*) as $k \to \infty$;*

(b) $\{E\|\hat{\theta}_k - \theta\|^2\}$ *converges as $k \to \infty$.*

Proposition 3. *Consider the problem of estimating μ and Σ completely.*

Let

$$\theta = [\mu' \vdots \sigma^{(1)'} \vdots \cdots \vdots \sigma^{(N)'}]' \tag{7a}$$

and

$$\hat{\theta}_k = [m_k' \vdots s^{(1)'} \vdots s^{(2)'} \vdots \cdots \vdots s^{(N)'}]', \tag{7b}$$

where

$$\sigma^{(i)} = [\sigma_{ii} \sigma_{i(i+1)} \cdots \sigma_{iN}]',$$

with $\sigma_{ij} = E(X_i X_j) - \mu_i \mu_j$, and

$$s^{(i)} = [s_{ii}^{(k)} s_{i(i+1)}^{(k)} \cdots s_{iN}^{(k)}]',$$

the s_{ij}'s being given by equation (4b).

If the conditions (C1), (C2) *and* (C3) *hold, then $\hat{\theta}_k$ (given by eqn.* (7b)*) converges with probability 1 to θ (given by eqn.* (7a)*).*

Proofs of Propositions

Before we give the formal proof of the propositions, we would like to make the following point.

If we subtract θ from both sides of the equation (1) we get, writing

$$\hat{\theta}_k^* = \hat{\theta}_k - \theta,$$

$$\hat{\theta}_k^* = \begin{cases} f(X_k) - \theta, & \text{for } k = 1, \\ \hat{\theta}_{k-1}^* - a_k Y_k^*, & \text{for } k > 1, \end{cases} \tag{8}$$

where

$$Y_k^* = \begin{cases} \hat{\theta}_{k-1}^* - (f(X_k) - \theta) \\ \quad \text{if } X_k \in G(m_{k-1}, \lambda_k), \\ 0 \quad \text{otherwise.} \end{cases} \tag{9}$$

This is because $0 < a_k \leq 1$ $\forall k$ by choice, so that for $k > 1$,

$$\hat{\theta}_k^* = \hat{\theta}_k - \theta = \hat{\theta}_{k-1} - a_k Y_k - \theta$$

$$= \begin{cases} \hat{\theta}_{k-1} - a_k(\hat{\theta}_{k-1} - f(X_k)) - \theta \\ \quad \text{if } X_k \in G(m_{k-1}, \lambda_k), \\ \hat{\theta}_{k-1} - \theta \quad \text{otherwise,} \end{cases}$$

$$= \begin{cases} (1-a_k)[\hat{\theta}_{k-1} - \theta] + a_k[f(X_k) - \theta] \\ \quad \text{if } X_k \in G(m_{k-1}, \lambda_k), \\ \hat{\theta}_{k-1}^* \quad \text{otherwise,} \end{cases}$$

writing $\theta = (1-a_k)\theta + a_k\theta$.

$$= \begin{cases} \hat{\theta}_{k-1}^* - a_k[\hat{\theta}_{k-1}^* - \{f(X_k) - \theta\}] \\ \quad \text{if } X_k \in G(m_{k-1}, \lambda_k), \\ \hat{\theta}_{k-1}^* \quad \text{otherwise.} \end{cases}$$

Thus Propositions 1 and 2 can be shown to be true if we can show that under the conditions assumed therein,

(i) $\hat{\theta}_k^* \to 0$ almost surely as $k \to \infty$;
(ii) $E[\|\theta_k^*\|^2]$ converges as $k \to \infty$,

To establish these we shall apply Theorems 1 and 2 directly, by showing that conditions (B1)–(B7) are true for $\hat{\theta}_k^*$ as defined by equations (8) and (9).

The conditions (B1) and (B2) are true because of our assumption (C1) and equations (8) and (9) respectively.

Proof of Proposition 1. Here,

$$E[Y_k^* | \hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_k^*]$$

$$= p_{k+1} E[\hat{\theta}_k^* - (X_{k+1} - \theta) | \hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_k^*]$$

(as $f(x) = x$ here)

$$= p_{k+1}[\hat{\theta}_k^* - E(x_{k+1} - \theta)]$$

(as X_{k+1} is independent of $X_1, X_2, \ldots, X_k$ and hence $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$)

$$= p_{k+1} \hat{\theta}_k^*$$

(as $\theta = \mu$ here and $E(X_{k+1}) = \mu$).

This verifies condition (B3), with

$$M_k(x) = p_{k+1} x, \quad x \in \mathbb{R}^N.$$

Further,

$$E[\|y_k^*\|^2 | \hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_k^*]$$

$$= p_{k+1} E[\|\hat{\theta}_k^* - (X_{k+1} - \theta)\|^2 | \hat{\theta}_1^*, \ldots, \hat{\theta}_k^*]$$

$$= p_{k+1}[\|\hat{\theta}_k^*\|^2 - 2\hat{\theta}_k^{*\prime} E(X_{k+1} - \theta) + E\|X_{k+1} - \theta\|^2]$$

(as X_{k+1} is independent of $\hat{\theta}_1^*, \ldots, \hat{\theta}_k^*$)

$$= p_{k+1}\left[\|\hat{\theta}_k^*\|^2 + E\sum_{n=1}^{N} (X_{(k+1)_n} - \mu_n)^2\right]$$

(as $E(X_{k+1} - \theta) = 0$)

$$= p_{k+1}\left[\|\hat{\theta}_k^*\|^2 + \sum_{n=1}^{N} \theta_n^2\right]$$

$$\leq \|\hat{\theta}_k^*\|^2 + \sum_{n=1}^{N} \sigma_n^2 \quad (\text{as } p_{k+1} \leq 1 \text{ and } a_k^2 \leq a_k \leq 1),$$

which means that condition (B4) holds with

$$a = \sum_{n=1}^{N} \sigma_n^2, \qquad b = 0, \qquad c = 1.$$

Also, $x' M_k(x) = p_{k+1} x' x \geq 0$ $\forall x \in \mathbb{R}^N$ which verifies condition (B5). That (B6) holds, is rather obvious, as

$$E\|f(X_1) - \theta\|^2 = E\|X_1 - \mu\|^2 = \sum_{n=1}^{N} \sigma_n^2 < \infty.$$

Finally, we can see that by virtue of our assumption (C2), the condition (B7) holds, for

$$\inf_{\eta < |x| < \eta^{-1}} [x' M_k(x)]$$

$$= \inf_{\eta < |x| < \eta^{-1}} [p_{k+1} x' x] > \delta\eta^2.$$

Thus Theorems 1 and 2 hold for $\hat{\theta}_k^*$. Hence the proposition is proved.

Proof of Proposition 2. Here

$$f(x) = \left[\underset{1\times N}{x^{(0)\prime}} \vdots \underset{1\times N}{x^{(1)\prime}} \vdots \underset{1\times(N-1)}{x^{(2)\prime}} \vdots \cdots \vdots \underset{1\times 1}{x^{(N)\prime}}\right]'$$

with

$$x^{(0)} = x,$$

$$\underset{q_i \times 1}{x^{(i)}} = [x_i^2, x_i x_{i+1}, \ldots, x_i x_N]',$$

$$q_i = N - i + 1.$$

Thus

$$E[Y_k^* \mid \theta_1^*, \theta_2^*, \ldots, \theta_k^*]$$
$$= E[\theta_k^* - (f(X_{k+1}) - \theta^*) \mid \theta_1^*, \theta_2^*, \ldots, \theta_k^*] p_{k+1}$$
$$= p_{k+1}[\theta_k^* - E\{f(x_{k+1}) - \theta\}]$$
(for the same reasons as before)
$$= p_{k+1}\theta_k^*, \quad (\text{as } E[f(X_{k+1})] = \theta).$$

This verifies (B3). Now,

$$E[\|Y_k^*\|^2 \mid \theta_1^*, \theta_2^*, \ldots, \theta_k^*]$$
$$= p_{k+1}[\|\theta_k^* - (f(X_{k+1}) - \theta)\|^2 \mid \theta_1^*, \theta_2^*, \ldots, \theta_k^*]$$
$$= p_{k+1}[\|\theta_k^*\|^2 - 2\theta^{*\prime}(E(f(X_{k+1}) - \theta)) + E\|f(X_{k+1}) - \theta\|^2] \quad (\text{as before})$$
$$= p_{k+1}[\|\theta_k^*\|^2 + E[\|f(X_{k+1}) - \theta\|^2]$$
(as $E[f(X_{k+1})] = \theta$).

However, as

$$E[\|f(X_{k+1})\|^2]$$
$$= E\left[\sum_{i=0}^{N} \|X_{k+1}^{(i)}\|^2\right] = \sum_{n=0}^{N} [E\|X_{k+1}^{(n)}\|^2]$$
$$= \sum_{n=1}^{N} E(X_n^2) + \sum_{j=1}^{N}\sum_{i\neq j}^{N} E(X_j^2 X_i^2)$$
$$\left(\text{since } E\|X_{k+1}^{(0)}\|^2 = E\left[\sum_{n=1}^{N} X_n^2\right]\right.$$
$$\left.\text{and } E\|X_{k+1}^{(i)}\|^2 = E\left[\sum_{i\neq j}^{N} X_i^2 X_j^2\right]\right)$$
$$= \sum_{n=1}^{N} (\sigma_n^2 + \mu_n^2) + \sum_{j=1}^{N}\sum_{i\neq j}^{N} \eta_{ij}$$
(by assumption (C3))
$$\leq K,$$

with K a finite positive constant independent of $\theta_1^*, \theta_2^*, \ldots, \theta_k^*$, we must have

$$E[\|Y_k^*\|^2 \mid \theta_1^*, \theta_2^*, \ldots, \theta_k^*]$$
$$\leq p_{k+1}[\|\theta_k^*\|^2 + K], \quad (\text{as } E[\|f(X_{k+1}) - \theta\|^2])$$
$$\leq \|\theta_k^*\|^2 + K$$
$$= E[\|f(X_{k+1})\|^2] - 2\theta' E(f(X_{k+1})) + \|\theta\|^2$$
$$= E[\|f(X_{k+1})\|^2] - \|\theta\|^2$$
(as $E[f(X_{k+1})] = \theta$)
$$\leq E\|f(X_{k+1})\|^2.$$

Thus condition (B4) is satisfied, with $a = K$, $b = 0$, $c = 1$. Further, conditions (B5), (B6) and (B7) also can seen to be true, as

$$x' M_k(x) = p_{k+1} x' x \geq 0, \quad \forall x \in \mathbb{P}^N,$$
$$E[\|f(X_1) - \theta_1^*\|^2] \leq E[f(X_1)]^2 < K < \infty,$$

as seen above, and

$$\inf_{\eta < |x| < \eta^{-1}} |x' M_k(x)| > \delta\eta^2 > 0,$$

because of our assumption (C2).

Thus Theorems 1 and 2 hold for $\{\theta_k^*\}$. This completes the proof of Proposition 2.

Proof of Proposition 3. The proof follows directly from Proposition 2 and the following lemma, if we note that $s_{ij}^{(k)}$ as given by eqn. (4b), is a continuous function, say, g_{ij} of θ_{k+1} (given by eqn (6a)), where $g_{ij}(x)$, $x \in \mathbb{R}^q$ is defined as

$$g_{ij}(x) = x_{r_{ij}} - x_i x_j$$

where

$$r_{ij} = \begin{cases} N + j & \text{if } i = 1, \\ N + \sum_{t=1}^{i-1} (N - t) + (j - i + 1) & \text{if } 1 < i \leq N, \end{cases}$$

and x_k denotes the k-th element of x, $k = 1(1)q$. Further, $g_{ij}(\theta) = \sigma_{ij}^* - \mu_i\mu_j = \sigma_{ij}$. Hence Proposition 3 follows.

Lemma. *Let $\{X_n\}$ be a sequence of random variables taking values in $\mathbb{R}^p$ and let $g : \mathbb{R}^p \to \mathbb{R}^q$ be a continuous map. Then $\{X_n\}$ converges with probability 1 to a ($\in \mathbb{R}^p$) implies that $g(X_n)$ converges with probability 1 to $g(a)$ ($\in \mathbb{R}^q$).*

References

Andrews, D.F. et al. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, NJ.

Chien, Y.T. (1970). The threshold effect of a nonlinear learning algorithm for pattern recognition. *Information Sciences* 2: 351–358.

Fu, K.S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning*. Academic Press, New York.

Huber, P.J. (1981). *Robust Statistics*. Wiley, New York.

Pal, S.K., A.K. Datta and D. Dutta Majumder (1980). A self-supervised and vowel recognition system. *Patt. Recogn.* 12, 27-34.

Schmetterer, L. (1968). Multidimensional stochastic approximation. In: P.R. Krishnaiah, Ed., *Multivariate Analysis-II: Proc. Second Int. Symp. Multivariate Anal.* Dayton, Ohio. Academic Press, New York.

Tsypkin, Ya.Z. (1973). *Foundations of the Theory of Learning Systems.* Academic Press, New York.