Consistent Set Estimation in k-Dimensions : An Efficient Approach

A. Ray Chaudhuri A. Basu S. K. Bhandari and **B. B. Chaudhuri**

INDIAN STATISTICAL INSTITUTE
203 B.T. ROAD CALCUTTA 700 035 INDIA

Abstract. Determining the shape of a point pattern is a problem of considerable practical interest and has applications in many branches of science related to Pattern Recognition. Set estimators of a nonparametric nature which may be used as shape descriptors should have several desirable properties. The estimators should be (a) consistent, i.e. Lebesgue measure of the symmetric difference of the actual region and the estimated should go to zero in probability; (b) computationally efficient; and (c) automatic, in the sense that the method should be able to detect the number of independent disjoint components in the region even when this number is unknown. None of the currently known estimators combine all these properties. A new shape descriptor called *s-shape* in the context of perceived border extraction of dot patterns in 2-D has been recently proposed. Here, a class of set estimators based on the s-shape is developed in k-dimensions, which combine all the above properties. These estimators are consistent not just under the uniform distribution, but also when samples are drawn under any continuous distribution. The order of error in estimation is independent of the dimensionality k. To illustrate the effectiveness of the proposed approach, a linear order algorithm readily derived from the definition is applied in digital domain. The role of δ that controls the structure of the estimator, is analyzed.

1. Introduction

From an early stage of human endeavor one problem of interest has been to find the shape of a point pattern. From astronomical studies to various application domains such as exploration of natural resources, urban planning, biomedical imaging, etc. the structural analysis of point set and shape recovery play an important role [3],[4]. In 2-D one can perceive the shape of the point set if they are clearly visible as well as fairly densely and more or less evenly distributed. Such a point set is referred to as a *regular dot pattern*. Intuitively speaking, the shape of a dot pattern is the bounded region that generates the pattern. Ray Chaudhuri et al. have introduced a shape descriptor called *s-shape* in the context of perceived border extraction of dot patterns in 2-D [6]. The idea behind the s-shape is as follows. Let the pattern plane be partitioned by a lattice of square grids of 'appropriate' length. Consider the union of grids containing points of the dot pattern. If the grid-length s is properly selected, this union (or a 'smooth' version of it) approximates the underlying region of the pattern.

Shape description may be viewed as an associated problem of the more basic question of set estimation from a finite number of sample points drawn from the set. One major problem of interest in set estimation is *consistency*. An estimator is *consistent* if it converges to the original set A as the number of points drawn from A tends to infinity (see Section 1.1). Here we extend the notion of s-shape and derive a class of set estimators in higher dimensions. The theoretical properties of these estimators are studied. The procedure is nonparametric in nature.

1.1 Consistent Estimation and Existing Results

Let $X_1, X_2, ..., X_n$ be k-dimensional independent and identically distributed (*i.i.d.*) observations from a distribution $\wp$ supported on a set $\alpha \subset \Re^k$. Let $cls(\alpha)$, $int(\alpha)$ and $\partial(\alpha)$ denote respectively the closure, interior and boundary of α; while λ, Δ and E are respectively the k-dimensional Lebesgue measure, symmetric difference operator and the expectation of a random variable.

Definition 1 : Let $\alpha^*_n \subset \Re^k$ be a set estimator of α based on $X_1, X_2, ..., X_n$. Then α^*_n is said to be a *consistent estimator* of α (denoted $\alpha^*_n \rightarrow \alpha$) if

$$\lim_{n \to \infty} E\left[\lambda\left(\alpha^*_n \Delta \alpha\right)\right] = 0. \tag{1}$$

Theorem I : Let $\alpha \subset \Re^2$ be a bounded Borel set whose boundary has Lebesgue measure 0. Let $<\varepsilon_n>$ be a sequence of positive numbers such that $n \rightarrow \infty$, $\varepsilon_n \rightarrow 0$ implies $n\,\varepsilon^2_n \rightarrow \infty$. Let $\alpha^*_n = \cup_{i=1}^n \left\{X \middle| \|X_i - X\| \le \varepsilon_n\right\}$. Then α^*_n is a consistent estimator of α under the assumption that $\wp$ is uniform.

The theorem is due to Grenander [1]. Note that there are infinitely many different sequences of such ε_n's with the required property. Since the choice of ε_n does not depend on $X_1, X_2, ..., X_n$, Grenander's class of estimators are not scale equivariant.

Another consistent set estimator based on the Minimum Spanning Tree (MST) is due to Murthy [5]. In this case, the radii ε_n's are made functions of $X_1, X_2, ..., X_n$ in the context of *compact regions*. A set α is said to be compact region if α is path-connected, compact, $cls(int(\alpha)) = \alpha$, and $\partial(\alpha)$ consists of finitely many rectifiable curves. Let φ_n denote the MST of $(X_1, X_2, ..., X_n)$.

Theorem II : $\alpha^*_n = \cup\left\{X \middle| \|Y - X\| \le h_n,\ Y \in \varphi_n\right\}$ is a consistent estimator of α where, $h_n = \sqrt{\frac{\varphi_n}{n}}$.

The above result is also true for any continuous distribution. However, the result can not be extended to the case of union of multiple compact regions unless the number of disjoint components is known. The above two theorems, established only in $\Re^2$, basically take the union of certain circular neighborhoods centering every sample point (in Theorem I) or points over the MST of sample points (in Theorem II) as an estimate of the original set α.

In Section 2, the s-shape in k-D and its derivative, a smooth version are formally defined. In Section 2.1, we establish the consistency of the s-shape under an uniform distribution in k-D. This result is then extended to general continuous distributions. In Section 2.2, the error rate in estimation is analyzed. One important result is that the order of error is independent of the dimensionality k. The consistency of the smooth s-shape is also established. In Section 3, an algorithm of linear order time complexity that is readily derivable from the definition of the proposed set estimators is applied to dot patterns illustrating its effectiveness in digital domain. The role of δ, the parameter controlling the structure of the estimators and the range of its values which appear to be intuitively and experimentally justified are analyzed. Finally, a general discussion is presented in Section 4.

2. The *s*-Shape and Its Derivatives as Set Estimators in *k*-D

Let $S_n \{X_1, X_2, ..., X_n\}$ be a set of n sample points in $\Re^k$. Let W_n be the *optimal* (with smallest k-volume) hyper-rectangle with boundary surfaces parallel to the $(k-1)$ dimensional coordinate planes of reference covering S_n, i.e. $S_n \subset W_n \subset \Re^k$. For a given hyper-cube (grid) of side-length s_n, let $\mathcal{F}(s_n)$ denote a lattice of grids on $\Re^k$, with sides parallel to the coordinate axes. For any grid g, let

$$\mathcal{G}(s_n) = \{g \mid g \cap W_n \neq \phi\}; \quad G(s_n) = \cup\{g \mid g \in \mathcal{G}(s_n)\} \tag{2}$$

$$\mathcal{H}(s_n) = \{g \mid g \cap S_n \neq \phi\}; \quad H(s_n) = \cup\{g \mid g \in \mathcal{H}(s_n)\} \tag{3}$$

Note that $G(s_n)$ is the set-union of grids over W_n while the *induced hull* $H(s_n)$ denotes the subset of $G(s_n)$ by picking those grids which contain at least one point. Let $\#H(s_n)$ denote the number of grids in $H(s_n)$. Then the Lebesgue measure of $H(s_n)$ in $\Re^k$ is $\lambda(H(s_n)) = \#H(s_n) \times (s_n)^k$.

Definition 2 : The induced hull $H(s_n)$ is called an *s-shape of* S_n.

By starting from the grid nearest to the center of reference of the coordinate axes, let the grids of $\mathcal{G}(s_n)$ be naturally (raster scan) ordered in a k-dimensional array. Then $\mathcal{G}(s_n)$ induces a k-D array (say,) $\prec z_{t_1,t_2,...,t_k} \succ$. In 2-D, its (t_1, t_2)-th element denotes the number of dots in the grid situated at t_1-th row, t_2-th column position.

Definition 3 : The *binary projection* of the above array is the array $\prec b_{t_1,t_2,...,t_k} \succ$, defined as

$$\left.\begin{aligned} b_{t_1,t_2,...,t_k} &= 1 \quad \text{if } z_{t_1,t_2,...,t_k} > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}\right\} \tag{4}$$

A set formed by the positions of non-zero entries in the binary projection, is referred to as an *object* and the rest is considered as *background*. A *hole* is a bounded part of the background embedded in the object. There is a one-to-one relation between the object and $\mathcal{H}(s_n)$. The object in the projection is also denoted by $\mathcal{H}(s_n)$.

Consider a 3×3×...×3 (k-tuple) array, (say,) τ having all entries equal to 1 as a k-D *structuring element* where the center of reference is located at the middle position of the array. *Binary closing* is a well known morphological filter which is defined as *dilation* followed by *erosion* [2]. It is an increasing, extensive and idempotent operator. Let the object be morphologically operated by a binary closing with τ in k-D integer space $\Im^k$. Let the resulting morphologically closed version of the object in the projection (with non-zero entries) be denoted by $\overline{\mathcal{H}}(s_n)$. Then,

$$\overline{\mathcal{H}}(s_n) = \left\{Z \in \Im^k \middle| t \in \Im^k,\ Z \in \tau_t \Rightarrow \mathcal{H}(s_n) \cap \tau_t \neq \Phi\right\}. \tag{5}$$

Let $\overline{H}(s_n)$ be the union of all grids whose corresponding $(t_1, t_2, ..., t_k)$-th positions in the projection are in $\overline{\mathcal{H}}(s_n)$. The closing 'smoothes' the set from outside : all background structures that can not contain the structuring element are added to the object. Thus, all holes of 'negligible size' are removed from $\overline{\mathcal{H}}(s_n)$, which is a superset of $\mathcal{H}(s_n)$. For example in 2-D, a grid $g \notin \mathcal{H}(s_n)$ having 5 non-empty grids among its 8-neighbors becomes a part of the closed version $\overline{\mathcal{H}}(s_n)$.

Definition 4 : $\overline{H}(s_n)$ is called the *smoothed induced hull* (*s-shape*) of the $H(s_n)$.

The consistency of $H(s_n)$ is first analyzed under an uniform distribution. Regarding the choice of s_n, a data driven procedure is proposed and the range where $H(s_n)$ remains consistent is established. The result is then generalized to the case of arbitrary continuous distributions.

2.1 Consistency of the *s*-Shape

Consider any set α, a finite union of connected subregions in $\Re^k$, each of which is bounded by a closed hyper-surface of finite area. Assume that the interior of α has a positive k-D volume (Lebesgue measure) but the boundary has Lebesgue measure zero. Let W be the optimal (in terms of k-D volume) hyper-rectangle with boundary surfaces parallel to the (k-1) dimensional coordinate planes of reference so that α lies in the interior of W. Without loss of generality let the volume of α be p (≤ 1), and that of W be 1. Let S_n $\{X_1, X_2, ..., X_n\}$ be the set of n points chosen randomly under a uniform distribution over α. Let the k-D volume of W_n covering S_n be $A_n \leq 1$. Let the side-length of hyper-cube grids in the lattice $\mathcal{F}(s_n)$ on $\Re^k$ be

$$s_n = n^{-\delta}\left(\sqrt[k]{A_n}\right) \qquad 0 < \delta < 1. \tag{6}$$

There are approximately $n^{k\delta}$ grids in the sublattice $\mathcal{G}(s_n)$ which consists of grids of $\mathcal{F}(s_n)$ intersecting W_n. For clarity of presentation, the following notations are used. Let T_n, I_n, B_n and H_n denote, respectively, the union of grids in $\mathcal{F}(s_n)$ intersecting α (some of them may not contain points of S_n), completely in interior of α, intersecting the boundary of α and the grids containing points of S_n. The latter set is the k-D *s*-shape. Let $\#T_n$, $\#I_n$, $\#B_n$ and $\#H_n$ denote the number of grids in the respective unions. Moreover, let n_I and n_B represent the number of points of S_n in I_n and B_n, respectively ($n_I + n_B = n$). If s_n is chosen as above, we shall show that the Lebesgue measure of the symmetric difference of H_n and α goes to 0 with probability 1 as n tends to ∞ for appropriate choices of δ. Now, by the *strong law of large numbers* (*SLLN*) any subregion of α with positive Lebesgue measure eventually has a point chosen from it in S_n with probability 1. Thus, as $n \to \infty$, $W_n \to W$ in the sense of (1), and $A_n \to 1$ with probability 1. Also,

$$\lambda(\alpha \cap T_n^c) \to 0. \tag{7}$$

where T_n^c denotes the complement of T_n.

We look at the proportion of empty grids among the $\#T_n$ grids intersecting the region α. If this proportion goes to 0 in probability, then by (7), $\lambda(\alpha \cap H_n^c) \to 0$ in probability. Since B_n approximates a (k-1) dimensional surface with Lebesgue measure zero (the boundary in question), and since the interior of α has positive Lebesgue measure (= p, the measure of α) we have

$$\lim_{n\to\infty} \lambda(B_n) = 0. \tag{8}$$

In fact,

$$\lim_{n\to\infty} \frac{\lambda(I_n)}{\lambda(T_n)} = 1 \quad \text{and} \quad \lim_{n\to\infty} \frac{\lambda(B_n)}{\lambda(T_n)} = 0. \tag{9}$$

Similarly,

$$\lim_{n\to\infty} \frac{n_I}{n} = 1 \quad \text{with probability one.} \tag{10}$$

Let $n_I \cong n\,a_n$ where $\lim_{n\to\infty} a_n = 1$. By the above results, it follows that

$$\lim_{n\to\infty} \frac{\#I_n}{n^{k\delta}} = p. \tag{11}$$

Suppose that n balls are thrown at random in $P_n = n^{\theta}/a$ boxes, $0 < \theta < k$, where a is a finite positive constant. Then the probability that any box remains empty is

$$\left(\frac{P_n - 1}{P_n}\right)^n = \left(1 - \frac{1}{P_n}\right)^n = \left(1 - \frac{a}{n^{\theta}}\right)^n, \tag{12}$$

which goes to zero in the limit for $\theta < 1$. Now, the expected proportion of empty cells among the $\#I_n$ cells in the interior of the region α is

$$\left(1 - \frac{1}{\#I_n}\right)^{na_n} = \left(1 - \frac{n^{k\delta}}{\#I_n}\frac{1}{n^{k\delta}}\right)^{na_n}. \tag{13}$$

Since in the limit the fraction ($n^{k\delta}/\#I_n$) tends to $1/p$, the limit of the expression in the right hand side of (13) becomes $e^{-\frac{1}{p}}$ if $\delta = 1/k$; equals to 0 if $\delta < 1/k$; and equals to 1 if $1/k < \delta < 1$. Thus, for any choice of $\delta < 1/k$, the expected proportion of empty cells among the squares completely in the interior of the region α goes to 0 for a sufficiently large n. As the proportion of empty grids is a nonnegative random variable, the proportion of empty cells also goes to 0 in probability by Markov's inequality. Also, by (9) as fraction of $\#B_n$ with $\#T_n$ is zero in the limit, the proportion of empty cells among $\#T_n$ is also zero in the limit.

Hence, $\lim_{n\to\infty} \frac{\#H_n}{\#T_n} = 1$ and H_n eventually covers α in probability. That is,

in probability as n goes to infinity, $\lambda(H_n^c \cap \alpha) \to 0$ (14)

Conversely, $\lambda(H_n \cap \alpha^c) \le \lambda((B_n \cup I_n) \cap \alpha^c) \le \lambda(B_n)$ (15)

Taking limit on both sides, $\lim_{n\to\infty} \lambda(H_n \cap \alpha^c) \le \lim_{n\to\infty} \lambda(B_n) = 0.$ (16)

Combining (14) and (16) it is established that $\lambda(H_n \Delta \alpha) \to 0$ in probability. As this symmetric difference is a bounded random variable, the next result follows.

Theorem III : Let S_n $\{X_1, X_2, \ldots, X_n\}$ be *i.i.d.* observations from a uniform distribution supported on a region α, a finite union of connected subregions in $\Re^k$, where each subregion is bounded by a closed hyper-surface of finite area. Let W_n be an optimal hyper-rectangle (in terms of hyper-volume) covering S_n with hyper-volume A_n. If $s_n = n^{-\delta}\left(\sqrt[k]{A_n}\right)$, $0 < \delta < 1/k$, then the s-shape $H(s_n)$ is a consistent estimator of α in k-dimensions.

In [7], the above result is generalized to the case of continuous distributions in 2-D. Combining it with Theorem III, the following theorem can be established.

Theorem IV : Let $X_1, X_2, \ldots, X_n$ be *i.i.d.* observations from a continuous distribution $\wp$ having support α on which its density function f is positive. Then the s-shape, $H(s_n)$ is a consistent estimator of α in k-dimensions under the conditions of theorem III.

2.2 Error Rate and Consistency of the Smooth *s*-Shape

The experimenter should have an idea of the order of error (the Lebesgue measure of the symmetric difference between α and α^*_n) when the procedure is terminated at a particular value of n and the estimate α^*_n has been determined. We provide an upper bound to this error when the points are drawn under a uniform distribution. We consider the hyper-cubes in the interior and the boundary of α separately.

The error in the interior E_I, related to the proportion of empty grids, is equal to

$$E_I = \lambda(I_n) \times \left(1 - c_1 \frac{1}{n^{k\delta}}\right)^{c_2 n} \quad \text{where } c_1 \text{ and } c_2 > 0. \quad (17)$$

By taking the logarithm of the R. H. S of (17), expanding, and exponentiating back, the leading term is found to be $c_3 e^{-n^{(1-k\delta)}}$ where c_3 is a positive constant. The error in the boundary E_B, satisfies

$$E_B \le \#B_n \times n^{-k\delta} \le \left(\frac{length\ of\ \partial(\alpha)}{n^{(k-1)\delta}}\right) \times n^{-k\delta} \le c_4 n^{-\delta} \quad (18)$$

where c_4 is a positive constant.

As the error in the boundary is dominant, the error in estimation is at most of order $O(n^{-\delta})$ and thus it is *independent of the dimensionality* of the data.

The smooth induced hull $\overline{H}(s_n)$, in general, is a better representation of the shape of a dot pattern than *s*-shape [6]. As $\overline{H}_n$ (which is an abbreviation of $\overline{H}(s_n)$) is a superset of $H(s_n)$ by (14), $\lambda(\overline{H}_n^c \cap \alpha) \to 0$ in probability. The boundary error may increase in case of $\overline{H}(s_n)$. But as $\lambda(H_n \cap \alpha^c) < 3^k \lambda(H_n \cap \alpha^c)$, $\lambda(\overline{H}_n \cap \alpha^c) \to 0$ in probability by (16). These results lead to the following theorem.

Theorem V : The smooth induced hull $\overline{H}(s_n)$ is also a consistent estimator of α under the conditions of Theorem IV.

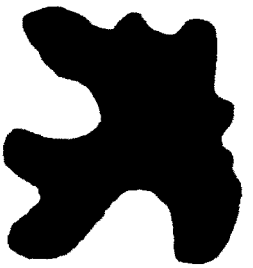

Fig.1 *A fish shaped region.*

3. Implementation in Digital Domain

To illustrate the effectiveness of our estimators, we have implemented the *s*-shape in two dimensional digital domain. The foreground in a binary digital image may be considered as α. Random samples of n object pixels are taken. The *area* of α, $\lambda(\alpha)$ is measured by the number of object pixels in α. An algorithm of linear order time complexity for computing the *s*-shape and its smooth version is readily derived from

the definition of binary projection [6]. It can be extended easily to k-D without altering the order of the algorithm.

Fig.2. *Dot patterns with sample size* n = 100, 1500, 3000.

The cardinality of α, the 'fish shaped' foreground in Fig.1 is 63903. Random samples from α with n = 100, 1500 and 3000 respectively, are shown in Fig.2. For δ = 0.45, their $H(s_n)$ and $\overline{H}(s_n)$ are presented in Fig.3(a)-(c) and Fig.3(d)-(f), respectively. Corresponding figures for δ = 0.49 are shown in Fig.3.(g)-(l). The ratio of $\lambda(\alpha^*_n \Delta \alpha)$ to $\lambda(\alpha)$ are plotted for $\alpha^*_n = H(s_n)$ and $\alpha^*_n = \overline{H}(s_n)$ against the sample size for δ = 0.45 and 0.49 in Fig.4. The asymptotic convergence of α^*_n is readily understood despite the limitation due to quantization. Fig.5 presents another example where our estimator is applied as a shape descriptor as well. The hole and disconnected components of the region (Fig.5(a)) are correctly recovered. Fig.5(b) and Fig.5(c) represent smooth s-shapes with δ = 0.49 and n = 1000, 2000 respectively.

3.1 Choice of δ

The choice of δ has considerable impact on the s-shape. For small values of δ, the boundaries of α^*_n are so crude that the s-shapes for $\delta \in (0, 0.45)$ appear to be of little practical utility. For larger values of δ, the s-shape has larger number of inconsistent holes. For δ = 0.5, the proportion of the area formed by such holes to the area of the region under estimation converges to a non-zero constant so that consistency fails. Thus 'smoothing' may be more useful for s-shapes with δ close to 0.5. For a given n, larger values of δ lead to small values of $\lambda(\alpha^*_n \cap \alpha^c)$ and smaller values of δ lead to small values of $\lambda((\alpha^*_n)^c \cap \alpha)$ i.e. values of δ near opposite ends of the allowable range are efficient in reducing complementary components of the symmetric difference.

On the whole, it appears that when single values of δ have to be recommended, then values around the center of the range [0.45, 0.5) or closer to 0.5 should be chosen (as larger values of δ reduce the dominant boundary error). When coupled with smoothing, values of δ close to 0.5 are recommended.

4. Discussion

The effectiveness of the s-shape as a set estimator has been illustrated in the examples, where it can be seen that the smooth s-shape can be viewed as a descriptor of the shape of the underlying region. Our proposed consistent set estimator is totally different from the consistent estimators reported earlier.

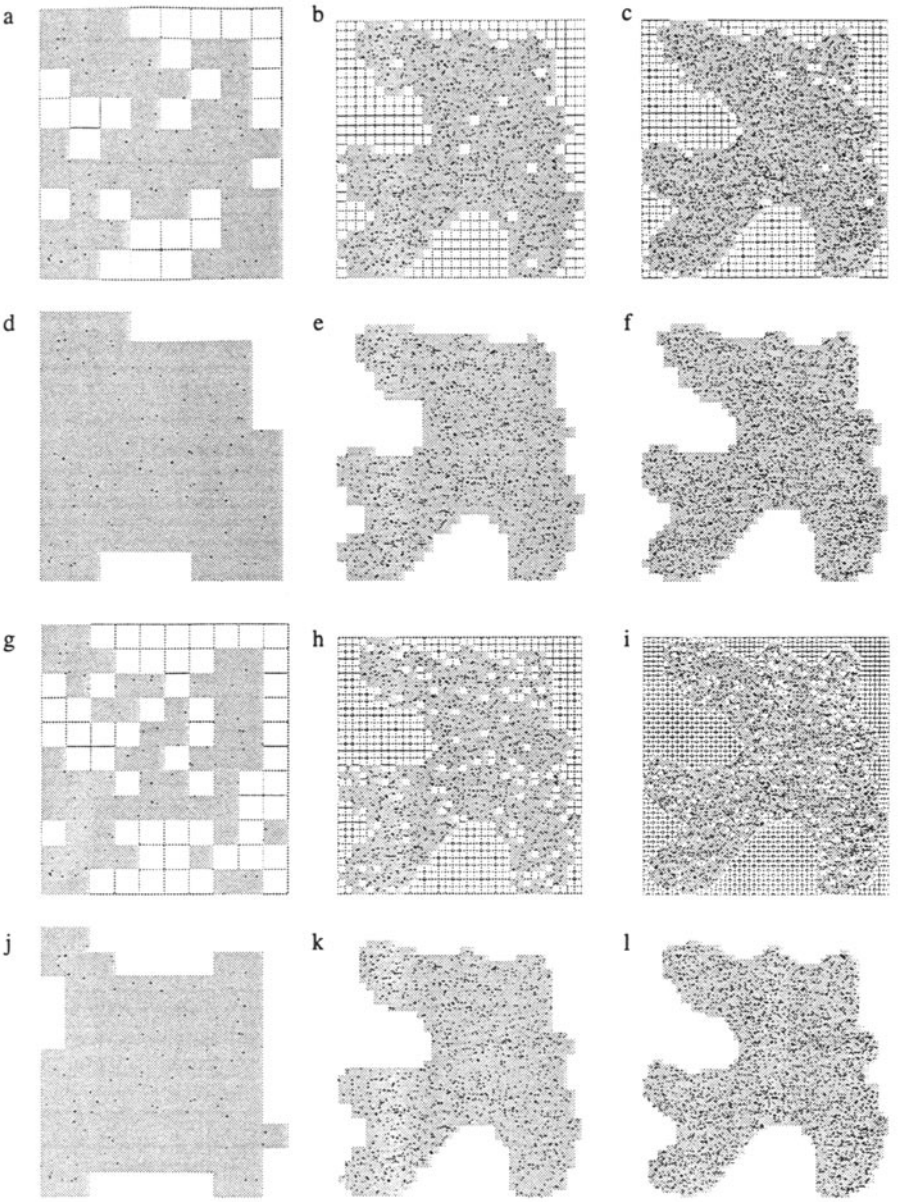

Fig.3 *s-shapes and their smooth versions with* $\delta = 0.45$ (a-f) *and* 0.49 (g-l) *respectively.*

The existing consistent estimators proposed by others are constructed by dilating either sample points itself or the edges of MST of the samples by a certain structuring disk. In our case, the optimal rectangular zone covering the sample set is partitioned by a lattice and the union of non empty grids is taken as α^*_n. One major advantage of our estimator is its computational efficiency. It is linear in terms of

cardinality of the sample i.e., for n observations the order of computational complexity is $O(n)$. On the other hand, in a MST based computable estimator, the construction of the MST alone takes $O(n \log n)$ provided sophisticated data structures are used. Thereafter, the MST has to be dilated by a structuring disk in the pattern space which will be quite involved in higher dimensions. Also, our estimator is fully unsupervised. The disconnected components are correctly detected as n increases. On the other hand, the MST approach needs prior knowledge about the number of such components in α.

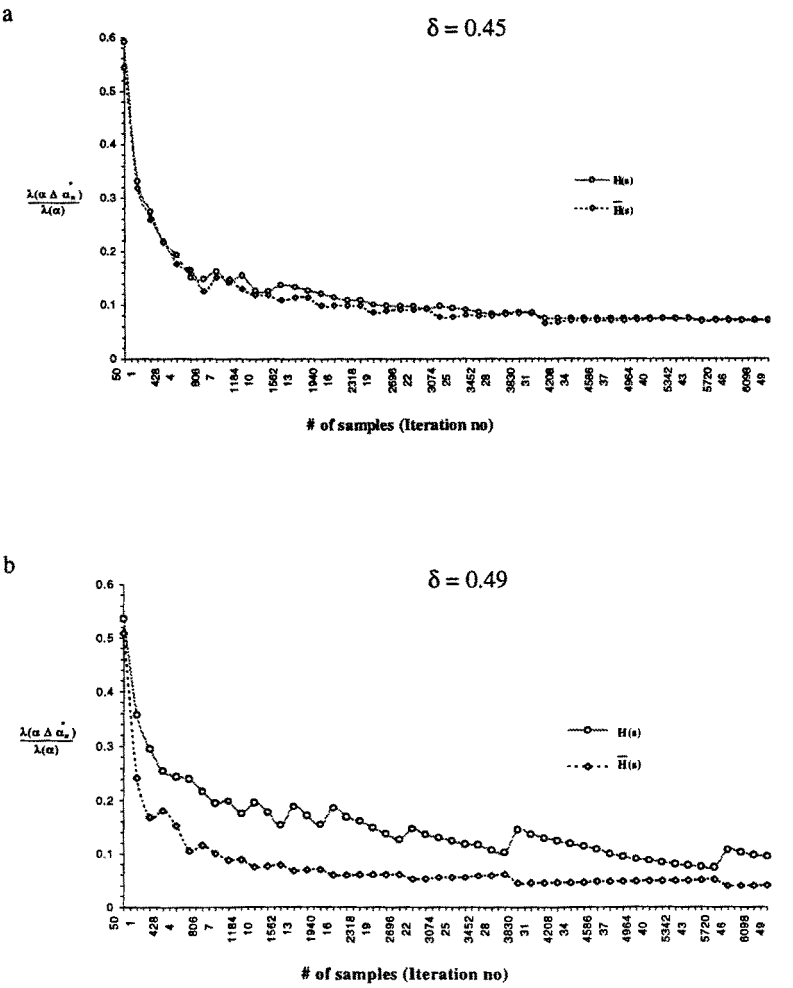

Fig.4 *Asymptotic convergence of* α^*_n

One needs to have an idea of the order of error when the procedure is terminated. We have provided an upper bound to this error (which is independent of the dimensionality) that may be used by practitioners as a guideline in determining a *stopping rule.* Note that such stopping rules are unavailable for existing estimators.

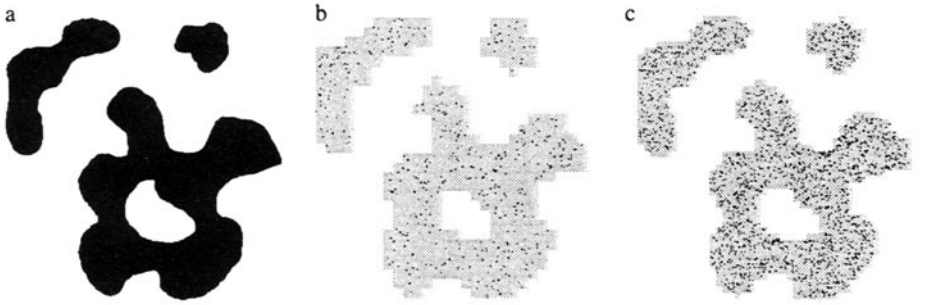

Fig.5 *The performance of the s-shape as a shape descriptor*

References

1. U. Grenander, *Abstract inference*, John Wiley, New York, 1975.
2. R. M. Haralick, S. R. Sternberg, and X. Zhuang, 'Image Analysis Using Mathematical Morphology', *IEEE Pattern Analysis and Machine Intelligence*,Vol. 9, pp. 523-550, 1987.
3. A. K. Jain, and R. Dubes, *Algorithms for clustering data*, Prentice Hall, New Jersy, 1988.
4. R. Laurini and D. Thompson, *Fundamental of spatial information systems*, The A.P.I.C. Series No. 37, Academic Press, London, 1992.
5. C. A. Murthy, '*On consistent estimation of classes in* $\Re^2$ *in the context of cluster analysis*', Ph. D. Thesis, Indian Statistical Institute, Calcutta, 1988.
6. A. Ray Chaudhuri, B. B. Chaudhuri and S. K. Parui, 'A novel approach to computation of the shape of dot pattern and extraction of its perceptual border', *Computer Vision and Image Understanding*, Vol.68, No.3, pp. 257-275, 1997.
7. A. Ray Chaudhuri, A. Basu, S. Bhandari and B. B. Chaudhuri, 'An efficient approach to consistent set estimation', *Tech. Rep.*, ASD/97/7, Applied Statistics Division, Indian Statistical Institute, Calcutta, India, August, 1997.