Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE @ DIRECT®

ELSEVIER Computer Vision and Image Understanding xxx (2003) xxx–xxx

Computer Vision and Image Understanding

www.elsevier.com/locate/cviu

# An efficient set estimator in high dimensions: consistency and applications to fast data visualization

A. Ray Chaudhuri,[a] A. Basu,[b] K. Tan,[c] S. Bhandari,[b] and B.B. Chaudhuri[b,*]

[a] *Computer Science and Engineering Department, Jadavpur University, Kolkata, West Bengal, India*
[b] *Indian Statistical Institute, 203 B.T. Road, Kolkata, West Bengal, India*
[c] *University of Illinois at Urbana Champaign Urbana, IL, USA*

Received 4 September 2002; accepted 17 October 2003

**Abstract**

Data visualization from a point set by estimating the underlying region is a problem of considerable practical interest and is an associated problem of set estimation. The most important issue in set estimation is consistency. Only a few existing point pattern shape descriptors that estimate the underlying region are consistent set estimators (a set estimator is consistent if it converges—in an appropriate sense—to the original set as the sample size increases). On the other hand, to be used as a shape descriptor, a set estimator should also satisfy several important criteria such as correct identification of number of components, robustness in the presence of noise and computational efficiency. Here we propose such a class of set estimators called s-shapes, which remain consistent in finite dimensions when the data are generated from any continuous distribution. These set estimators can be easily computed and effectively used for fast data visualization. Detailed studies on their performance such as error rates, robustness in presence of noise, run-time analysis, etc., are also performed.

*Keywords:* Set estimation; Consistency; Point set; Shape descriptors; Visualization; High-dimensional data

[*] Corresponding author. Fax: +91-33-2577-6680/3035.
*E-mail address:* bbc@isical.ac.in (B.B. Chaudhuri).

1077-3142/$ - see front matter © 2003 Published by Elsevier Inc.
doi:10.1016/j.cviu.2003.10.002

## 1. Introduction

Finding the underlying region of a point set is a theoretical problem with substantial practical relevance. In 3-dimensions (3D), data visualization by reconstructing the volume from representative points has several applications [1–4]. In biomedical imaging, 3D reconstruction of jell-like tissue-mass from stained particles is often necessary in diagnosis. Among others, geographical information system (e.g., in case of synthetic aperture radar images, the underlying region of scattering centers can effectively characterize the target object), space science (estimating shape and size of remote galaxy), atmospheric study (analyzing cloud mass), etc., are some relevant areas where shape or region estimation from a point set is a very important problem.

This shape description from point set may be viewed as an associated problem of the more basic question of *set estimation* from a finite number of sample observations drawn from the set [5,6]. The boundary of the estimated set can then be used as the shape descriptor. For the purpose of set estimation itself, no shape-topology based information is necessary and the whole computation should be unsupervised and based on set theoretic formalisms. For a set estimator, the most important issue is its *consistency*. A set estimator is consistent if it converges (in an appropriate sense defined in the following subsection) to the original set, as the number of observations drawn from it becomes large. In the literature, there exist several shape (boundary) descriptors of point-patterns. Only a few of them have attempted to establish consistency of their proposed descriptors as set estimators. The estimator that might be used as a shape descriptor should satisfy a few important criteria. The estimator should (a) automatically detect the number of independent disjoint components in the true region; (b) be robust in presence of *additive noise* (observations outside the region of interest that may be added at the time of data acquisition), and (c) be computationally efficient. None of the set estimators known to us that may be used as shape descriptors of point-patterns combine all the above properties.

Ray Chaudhuri et al. [7] introduced a measure called *s-shape hull* (or simply *s-shape*) in the context of perceived border extraction of point sets in 2D. The idea behind the s-shape is as follows: let the pattern plane be partitioned by a lattice of square cells of ‘appropriate’ length, $s$. Consider the hull, which is the union of the cells containing points of the dot pattern. If the cell-length $s$ is properly selected, this hull or a ‘smooth’ version of it approximates the underlying region of the point set or *the region of support*. However, that shape descriptor was not consistent as a set estimator. Later it was demonstrated that the particular choice of cell-size used in [7] led to inconsistent set estimation and a more judicious selection of the cell-size was required [8].

In this paper, the notion of s-shape is extended to derive a class of consistent set estimators in the context of high-dimensional data by appropriately modifying the selection criterion for $s$. The spirit of the procedure is non-parametric in nature. Smooth versions of s-shape are derived by binary morphological transformations. The *serial compositional properties* [9] of basic morphological operators as well as *conditional erosion* [10] are utilized for fast computation. The notion of *s-shape spectrum*, useful for a given set of observations (fixed $n$) is also presented. The set-consistency of s-shape and its derivatives are established not only when the observations are generated by a uniform distribution over the region of support but for any continuous distribution. The error analysis reveals that the order of error is independent of the dimensionality of the data. The run-time of s-shape computation is compared with another competing consistent set estimator.

Another important aspect of the paper is data-visualization via s-shape. A simple algorithm with linear time complexity for s-shape computation, valid for any finite dimension is presented. In particular, in 2D data, the robustness of s-shape in the presence of noise is tested. Volume visualization in 3D via s-shape demonstrates the ability of the estimator to distinguish multi-components, even when one component is completely embedded in another. From the runtime analysis it is apparent that s-shape based visualizer is considerably fast.

We begin the following section by formally defining consistency of set estimators. Some important results existing in literature on this topic are also mentioned. It is followed by the organization of the paper.

### 1.1. Consistent set estimation and existing results

Let $x_1, \ldots, x_n$ be the realization of $n$ independent and identically distributed (i.i.d.) $d$-dimensional ($d$-D) observations drawn from a distribution $\wp$ which is supported on a set $\alpha$, a finite union of bounded and connected subregions in the $d$-dimensional real space $\Re^d$. Let $\text{cls}(\alpha)$, $\text{int}(\alpha)$, and $\partial(\alpha)$, respectively, denote the closure, interior and boundary of $\alpha$. Let $\alpha_n^* \subset \Re^d$ be a set estimator of $\alpha$ based on the random sample $x_1, \ldots, x_n$.

**Definition 1.** $\alpha_n^*$ is a *consistent set estimator* of $\alpha$ (denoted as $\alpha_n^* \to \alpha$) if the expectation of the $d$-D volume of the symmetric difference between $\alpha_n^*$ and $\alpha$ tends to zero as $n \to \propto$. That is,

$$\lim_{n\to\infty} E[\lambda(\alpha_n^* \Delta \alpha)] = 0, \tag{1}$$

where $\lambda(A)$ denotes the $d$-D volume of a set $A$ in $\Re^d$. A more generalized definition and treatment can be found in [5] where Grenander proposed a consistent set estimator for data in real plane via the following theorem.

**Theorem 1.** *Let $\alpha \subset \Re^2$ be a bounded set where $\lambda(\partial(\alpha)) = 0$ and $\langle \varepsilon_n \rangle$ be a sequence of positive numbers such that $n \to \propto$, $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \propto$. Let $\alpha_n^* = \bigcup_{i=1}^n \{X \| \|x_i - X\| \leqslant \varepsilon_n\}$. Then $\alpha_n^*$ is a consistent estimator of $\alpha$ under the assumption that $\wp$ is uniform.*

The same problem is considered by Mandal et al. [11] where the circular disk surrounding each $x_i$ is replaced by a rectangular neighborhood. Since the choice of $\varepsilon_n$ in Theorem 1 does not depend on $x_1, \ldots, x_n$, Grenander's class of estimators does not have the scale equivariance property and thus lacks a very important desirable feature.

Another consistent set estimator based on the Minimum Spanning Tree (MST) is due to Murthy [12]. In this case, the radii $\varepsilon_n$'s are made functions of $x_1, \ldots, x_n$ in the context of *compact regions* (see [13] for definitions).

**Theorem 2.** *Let* $\alpha_n^* = \cup\{X | \|Y - X\| \leqslant h_n,\ Y \in \varphi_n\}$ *where* $h_n = \sqrt{l_n/n}$ *and* $l_n$ *is the length of the MST* $\varphi_n$*. Then* $\alpha_n^*$ *is a consistent estimator of* $\alpha$.

The result also holds when $\wp$ is any continuous distribution. However, the result cannot be extended to the case of union of multiple disjoint compact regions unless the number of disjoint components is known. This method is sensitive to additive noise.

The above two theorems, established only on $\Re^2$, basically take the union of certain circular neighborhoods centering every sample point (in Theorem 1) or points over the MST of sample points (in Theorem 2) as an estimate of the original set $\alpha$.

In the context of shape description, there are a few methods on the boundary shape computation of a point set by a triangular mesh derived from Voronoi/Delaunay tessellations Two classical works on this field are $\alpha$-shape proposed by Edelsbruner et al. [14] and for perceptual shape recovery from point set by Ahuja and Tceryan [15]. Worring and Smeulders [16] considered the set consistency of $\alpha$-graph, a variant of the compact region bounded by the $\alpha$-shape, in 2D. They established that the $\alpha$-graph of a connected set converges to itself. This is akin to establishing the *Fisher consistency* in the context of parametric statistical estimation [17], rather than showing strong consistency as a limiting result of the sample size. Recently, Amenta et al. [18] have given an elegant definition called, *crust* for surface reconstruction. It guarantees that for a "good sample" (having an appropriate sampling density depending on the local surface curvature) from a smooth surface the reconstructed surface will be topologically correct and convergent to the original surface as the sampling density increases. However, the method could not be directly extended to the case where observations came from the interior of the object rather than from its surface alone. In addition, the result is sensitive to noisy data. Two other recent papers for 3D surface generation from Voronoi/Delaunay tessellations, by Attali [19] and by Melkami et al. [20] may be mentioned in this connection.

As mentioned earlier, s-shape as a consistent set estimator was introduced in 2D by Chaudhuri et al. [8]. Some initial indication about the possible extensions of the method for high-dimensional data are available in [21].

### *1.2. Organization of the paper*

In Section 2, *d*-dimensional s-shape based set estimators and their smooth versions are defined. The basic cell-size estimation criterion is described and the notion of *s-shape spectrum*, useful for a given set of observations (fixed *n*) is presented. Section 3 deals with set-consistency of the s-shape. Sections 3.1 and 3.2, establish consistency under uniform and general continuous distributions, respectively. Set consistency of the smooth version of s-shape, obtained through some morphological transforms is considered in Section 3.3. The s-shape spectrum and its set consistency are also presented. In Section 3.4, the error rate in estimation is analyzed.

The details of computer implementation of s-shape and its smooth versions are presented in Section 4. It begins with the s-shape computation algorithm (Section 4.1). In Section 4.2, smoothing of s-shape via morphological operators is described.

In Section 5, data visualization by s-shape and other experimental results are demonstrated. It starts with results in 2D data. The rate of convergence and comparison with MST based estimator as well as performances in presence of noise are demonstrated. Section 5.1 discusses the role of $\delta$, the parameter controlling the structure of the estimators. In Section 5.2, s-shape as a 3D volume visualizer is presented which includes the run-time analysis. Section 6 presents a summary and directions for future work.

## 2. Proposed class of set estimators (s-shape) and their geometrical interpretations

Let $\Upsilon_n = \{x_1, \ldots, x_n\}$ be a random sample of size $n$ drawn from the support set $\alpha$ where $\alpha$ is a finite union of bounded and connected sub-regions in the $d$-dimensional real space $\mathfrak{R}^d$. Let $W_n$ be the isothetic (boundary surfaces perpendicular to the $(d-1)$-dimensional coordinate planes of reference) hyper-rectangle with the smallest $d$-dimensional volume covering $\Upsilon_n$, i.e., $\Upsilon_n \subset W_n \subset \mathfrak{R}^d$. For hyper-cubes (cells) of side-length $s$, let us consider a lattice of isothetic cells $g$ on $\mathfrak{R}^d$, with surfaces parallel to the $(d-1)$-dimensional coordinate planes. For any such lattice, define

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

$$\mathcal{H}(s_n) = \{g \,|\, g \cap \Upsilon_n \neq \phi\};\;\; H(s_n) = \cup\{g \,|\, g \in \mathcal{H}(s_n)\}. \tag{3}$$

Here $G(s_n)$ denotes the union of cells intersecting $W_n$.

**Definition 2.** The *$d$-dimensional s-shape $H(s_n)$* is the subset of $G(S_n)$ obtained by joining the cells which contain at least one point from$\Upsilon_n$. $\mathcal{H}(s_n)$ is called the *lattice representation of the s-shape $H(s_n)$.*

The $d$-dimensional volume of $H(s_n)$ can be given by $\lambda(H(s_n)) = \#\mathcal{H}(s_n) \times (s_n)^d$ where $\#\mathcal{H}(s_n)$ is the number of cells in $\mathcal{H}(s_n)$.

Starting from the cell nearest to the center of reference of the coordinate axes, let the cells of $\mathcal{G}(s_n)$ be ordered in a raster fashion in a $d$-dimensional array. For example, for 2D case, starting from the cell in first column and first row one has to move along the row (in the direction of dimension 1) by crossing the columns till the last column attains. Then along the first column move one step down (in the direction of dimension 2) and resume moving along the second row and so on. Then $\mathcal{G}(s_n)$ induces a $d$-dimensional array (say,) $\prec z_{t_1,\ldots,t_d} \succ$ which indicates the number of points in the cell at position $(t_1, \ldots, t_d)$. Consider the binary projection of the array $\prec z_{t_1,\ldots,t_d} \succ$, say $\prec b_{t_1,\ldots,t_d} \succ$, where $b_{t_1,\ldots,t_d}$ is one or zero as $z_{t_1,\ldots,t_d}$ is positive or otherwise. From the geometric point of view, the set consisting of the positions of non-zero entries (grids) in that binary projection may be considered as the *foreground* (*object*) while the rest is the *background*. Because of one-to-one correspondences, one can interchangeably use $\mathcal{G}(s_n)$ for $\prec b_{t_1,\ldots,t_d} \succ$ and $\mathcal{H}(s_n)$ for the foreground, respectively.

By generalizing the definition of neighbors in a 2D digital image, connectivity in cells of $\mathcal{G}(s_n)$ in $\mathfrak{R}^d$ can be defined [22]: Any two cells in the foreground are *neighbors* if they meet at a point, line, or a hyper-plane of dimension less than $d$ in $\mathfrak{R}^d$. In case of background, two cells are neighbors only if they meet at a hyper-plane of dimension $d-1$. Two cells $g_1$, $g_l$ of same type (both in foreground or both in background) in $\mathcal{G}(s_n)$ in $\mathfrak{R}^d$ are *connected by a path* of cells (in object or background exclusively) in the form of a sequence, say, $g_1, \ldots, g_i, \ldots, g_l$ so that $g_{i-1}$ is a neighbor of $g_i$ in $\mathcal{G}(s_n)$. A *component* in $\mathcal{H}(s_n)$ is the subset of cells where each cell in it is connected by a path to any other cell in the same. The notion of *hole* in this context can be defined as a connected component in the background with finite volume. Geometrically, it is a union of empty cells completely surrounded by components of the s-shape.

### 2.1. Smoothed s-shape

It is apparent that the original s-shape may be corrupted with inconsistent small holes and unwanted border-cracks particularly when cell size is quite small. Thus, some smoothing that can remove such holes and cracks from the s-shape will be useful.

Let $\overline{\mathcal{H}}(s_n)$ denote a superset of $\mathcal{H}(s_n)$ and is defined as follows:

$$\overline{\mathcal{H}}(s_n) = \{J \in \mathfrak{I}^d \,|\, t \in \mathfrak{I}^d,\ J \in \tau_t^d \Rightarrow \mathcal{H}(s_n) \cap \tau_t^d \neq \phi\}, \tag{4}$$

where, $\tau^d$ is a $3 \times 3 \times \cdots \times 3$ ($d$-tuple) array with the center of reference located at the middle position and $\tau_t^d$ denotes the translation of $\tau^d$ to $t$.

It can be shown that $\overline{\mathcal{H}}(s_n)$ is the *binary morphological closing* of $\mathcal{H}(s_n)$ with $\tau^d$ as the *structuring element* in $d$-D integer space $\mathfrak{I}^d$. The binary closing is a well-known morphological filter that is defined as dilation followed by erosion with the same structuring element [23,24]. Note that binary dilation and erosion of a discrete set $X$ by a (symmetric) structuring element $Y$ is, respectively, defined by $X \oplus Y = \bigcup_{y \in Y} X_y$ and $X \Theta Y = \bigcap_{y \in Y} X_y$, where $X_y$ denotes the translation of $X$ to $y$. Closing 'smoothes' the set from outside. Holes and outside narrow cleavages of 'negligible-size' (less than size of $\tau^d$) are filled up and become part of the object. For example in 2D, a background grid (pixel) $g$ in $\mathfrak{I}^2$ having 5 foreground grids in its 8-neighbors with any configuration becomes a foreground pixel in the closed version.

However, it is often the case that some noise is added at the time of data (sample) acquisition. To take care of such additive noisy data, the dual of closing namely *opening*, i.e., the erosion followed by the dilation with same structuring element may be considered. If one applies opening directly on $\mathcal{H}(s_n)$ then the resultant set $\overset{\circ}{\mathcal{H}}(s_n) = \{\cup \tau_t^d \,|\, t \in \mathfrak{I}^d,\ \tau_t^d \subset \mathcal{H}(s_n)\}$ preserves only those parts where the structuring element $\tau^d$ can be placed completely inside $\mathcal{H}(s_n)$ and rest will be removed from the set. Due to presence of possibly several small holes and cracks in $\mathcal{H}(s_n)$, any opening is only effective when such small holes and cracks are already taken care. Thus, in presence of additive noise, $\overset{\circ}{\overline{\mathcal{H}}}(s_n)$, the morphologically *clopen transform* (closing followed by opening with the same structuring element) of $\mathcal{H}(s_n)$ is taken as the smooth version.

Let $\overline{H}(s_n)$ and $\overset{\circ}{\overline{H}}(s_n)$ denote the unions of all cells in $\mathfrak{R}^d$ whose corresponding reference positions belong to $\overline{\mathcal{H}}(s_n)$ and $\overset{\circ}{\overline{\mathcal{H}}}(s_n)$, respectively.

**Definition 3.** The induced hull $\overline{H}(s_n)(\overset{\circ}{\overline{H}}(s_n))$ is the (*default*) *smooth version of the s-shape* $H(s_n)$ when all observations are (not necessarily) from the support set $\alpha$.

### 2.2. Choice of cell-size and the s-shape spectrum

The most crucial factor in computation of the s-shape is the estimation of the side-length $s_n$ of the cells. Let us assume that for a set of $n$ observations, the $d$-dimensional volume of isothetic optimal covering rectangle $W_n$ be $V_n$. Then the side-length $s_n$ is chosen (as a function of a single parameter $\delta$) as

$$s_n = n^{-\delta}\left(\sqrt[d]{V_n}\right), \quad 0 < \delta < 1. \tag{5}$$

To make the class of s-shape based set estimators more robust for a given sample (fixed $n$), we introduce the notion of s-shape spectrum in $d$-dimensions. It consists of successive finer (with smaller cell size) approximations of the s-shape from the previous one. When two adjacent s-shapes become close enough under certain relative measure, the finer one is selected as the eventual candidate. This way the abrupt increase of the cell size of s-shape due to few observations on a large support set or due to presence of additive noise can be adjusted. In case of noisy data, as discussed above, cells covering scattered noisy points in the s-shape are removed by the morphological clopening.

The shape spectrum is formally defined as follows:

Consider the sequence $\langle H(s_n^t)\rangle$, $t \in \mathfrak{I}$, where

$$s_n^t = n^{-\delta}\sqrt[d]{\lambda(H(s_n^{t-1}))}, \quad \text{and} \quad H(s_n^0) = W_n. \tag{6}$$

Thus, $\lambda(H(s_n^0)) = V_n$ and $s_n^1 = s_n$.

If one takes $\delta = 1/d$ then it can be shown that $\langle H(s_n^t)\rangle$ is a sequence where cell-size is gradually decreasing and it converges after finite steps. In that situation, the sequence converges to a hull with a maximal volume where each non-empty cell contains only one observation. However, when $\delta < 1/d$ the strict monotonicity may not be preserved. More discussion on the choice of $\delta$ is available in Section 5.1.

**Definition 4.** The (finite) sub-sequence of $\langle H(s_n^t)\rangle$ starting from the first element satisfying monotonic decreasing criterion is the *s-shape spectrum* of the given data.

## 3. Consistency of s-shape based set estimators

In this section, the consistency of the s-shape $H(s_n)$ is first analyzed under a uniform distribution. A data driven procedure is proposed regarding the choice of the cell-size for which $H(s_n)$ remains consistent is established. Later it is generalized to the case where the sample is generated randomly from any continuous distribution.

### 3.1. Points from a uniform distribution

We assume here that $\wp$, the distribution from which $n$ i.i.d. observations $\Upsilon_n = \{x_1, \ldots, x_n\}$ are drawn, is uniform. Recall that $\wp$ is supported $\alpha$, a finite union of bounded and connected subregions in the $d$-dimensional real space $\Re^d$. Without loss of generality, let the ($d$-D) volume of $\alpha$ be $p$ $(0 \leqslant p \leqslant 1)$ and that of $W$, the optimal isothetic hyper-rectangle covering $\alpha$ be 1. As mentioned earlier, let the ($d$-D) volume of the optimal hyper-rectangle $W_n$ covering $\Upsilon_n$ be $V_n$ ($\leqslant 1$) i.e., $\lambda(W_n) = V_n$ and $s_n = n^{-\delta}(\sqrt[d]{V_n}),\ 0 < \delta < 1$.

Henceforth, unless noted otherwise, $\mathcal{G}(s_n)$, $\mathcal{H}(s_n)$, $G(s_n)$, $H(s_n)$ will be simply denoted by $\mathcal{G}_n$, $\mathcal{H}_n$, $G_n$, and $H_n$, respectively. Let $T_n$, $I_n$, $B_n$ denote, respectively, the union of cells in the lattice in $\Re^d$ intersecting $\alpha$ (some of them may not contain points of $\Upsilon_n$), completely in the interior of $\alpha$, intersecting the boundary of $\alpha$. Let $n_I$ and $n_B$ represent the number of points of $\Upsilon_n$ in $I_n$ and in $B_n$, respectively. Clearly, $T_n = I_n \cup B_n$ and $n = n_I + n_B$.

By the Strong law of large numbers (SLLN) any sub-region of $\alpha$ eventually has a point chosen from it with probability 1 as $n \to \infty$. Thus $W_n \to W$ in the sense of (1) as $n \to \infty$ and $V_n \to 1$ with probability 1. Note, $\#G_n$ is approximately $n^{d\delta}$, and $\lim_{n\to\infty}(n^{d\delta}/\#G_n) = 1$. Since the boundary has ($d$-D) volume zero, it follows that

$$\lim_{n\to\infty} \frac{\#I_n}{\#G_n} = \lim_{n\to\infty} \frac{\#I_n}{n^{d\delta}} = p, \tag{7}$$

while

$$\lim_{n\to\infty} \frac{\#B_n}{\#T_n} = 0 \quad \text{and}_{n\to\infty} \frac{\#n_I}{\#n} = 1.$$

Thus, one can use the representation $n_I \cong na_n$ where $a_n \leqslant 1$ and $\lim_{n\to\infty} a_n = 1$. By a simple probability argument, under the assumption that the observations are i.i.d., the expected proportion of empty cells (cells not containing points of $\Upsilon_n$) among the $\#I_n$ cells in the interior of $\alpha$ is

$$\left(1 - \frac{1}{\#I_n}\right)^{na_n} = \left(1 - \frac{n^{d\delta}}{\#I_n}\frac{1}{n^{d\delta}}\right)^{na_n}. \tag{8}$$

By (7), the limit of the expression as $n \to \infty$ on the right-hand side of the above equals to $e^{-1/p}$ if $\delta = 1/d$; equals zero if $\delta < 1/d$; and equals one if $1/d < \delta < 1$.

Thus, the expected proportion of empty cells in the interior of the region $\alpha$ goes to zero whenever $\delta < 1/d$. Since the proportion of empty cells in the interior of $\alpha$ is a non-negative random variable, this proportion itself goes to zero in probability as $n \to \infty$ by Markov's inequality. Also, since $\lim_{n\to\infty}(\#B_n/\#T_n) = 0$ (where $\lambda(B_n) \to 0$ as $n \to \infty$) it follows that proportion of empty cells among $T_n$ goes to zero in probability. Since $H_n$ is the union of non-empty cells, $\lambda(H_n^c \cap \alpha) \to 0$ in probability.

On the other hand, $(H_n \cap \alpha^c) \subset (I_n \cup B_n) \cap \alpha^c = (I_n \cap \alpha^c) \cup (B_n \cap \alpha^c) = (B_n \cap \alpha^c) \subset B_n$. Thus, $\lambda(H_n \cap \alpha^c) \leqslant \lambda(B_n)$. Since, $\lambda(B_n) \to 0$ as $n \to \infty, \lambda(H_n \cap \alpha^c) \to 0$ as $n \to \infty$.

Combining the above two results, $\lambda(H_n \Delta \alpha) \to 0$ in probability. In addition, since $\lambda(H_n \Delta \alpha)$ is a bounded random variable, $E(\lambda(H_n \Delta \alpha)) \to 0$ as $n \to \infty$.

**Theorem 3.** *Let $\Upsilon_n = \{x_1, .., x_n\}$ be i.i.d. observations from a uniform distribution supported on $\alpha$, a finite union of bounded connected subregions in $\Re^d$. Let $W_n$ be the optimal hyper-rectangle with volume $V_n$ covering $\Upsilon_n$ in d-dimensions. If $s_n = n^{-\delta}\left(\sqrt[d]{V_n}\right)$, $0 < \delta < 1/d$, then the s-shape $H(s_n)$ is a consistent estimator of $\alpha$ in d-dimensions.*

The positive fraction $\delta$, the only parameter used in the s-shape description, acts like a resolution parameter. A value of $\delta$ closer to zero means larger-sized cell and the resulting s-shape is a much cruder representation compared to the case where $\delta$ is close to $1/d$. More discussion on choice $\delta$ in different situations is in Section 5.1.

### 3.2. Points from arbitrary continuous distributions

Let $\alpha$ be as defined in Section 3.1 where $\lambda(\alpha) = p$, $(0 < p \leqslant 1)$ and $f(> 0)$ be a continuous density function supported on $\alpha$. Let $\Upsilon_n$ be a set of points drawn at random from $\alpha$ according to $f$. $W$ and $W_n$ are defined as in Section 3.1. with $\lambda(W) = 1$ and $\lambda(W_n) = V_n$.

Let $\wp(Q)$ be the probability of $Q \subset \alpha$ under $f$. Given any $\varepsilon > 0$, one can choose a $m$ sufficiently large so that the region $\alpha_m = \{x \mid x \in \alpha, \frac{1}{m} < f(x) < m\}$ satisfies $\lambda(\alpha_m) > p - (\varepsilon/2)$. Let $\wp(\alpha_m) = p'$. Also assume that $\lambda(\partial(\alpha_m)) = 0$. Let $T_{m,n}$, $I_{m,n}$, $H_{m,n}$, denote, respectively, the union of cells in the lattice intersecting $\alpha_m$, completely in the interior of $\alpha_m$, and the cells in $\alpha_m$ containing points of $\Upsilon_n$. Rest of the notations are identical to the previous section. As in the previous case, $W_n \to W$ and $V_n \to 1$ in probability.

Notice that the definition of $\alpha_m$ implies, for any cell $g$ in $I_{m,n}$ and a point $z$ in $\Upsilon_n$, $\frac{1}{m}\lambda(g) < \wp(z \in g) < m\lambda(g)$ and $\wp(z \in I_{m,n}) < m\#I_{m,n}\lambda(g)$.

Thus,

$$\wp(z \in g | z \in I_{m,n}) = \frac{\wp(z \in g)}{\wp(z \in I_{m,n})} > \frac{\frac{1}{m}\lambda(g)}{m\#I_{m,n}\lambda(g)} = \frac{1}{m^2\#I_{m,n}}. \tag{9}$$

Let $m_I$ denote the number of points in the interior of $\alpha_m$. We use the representation $m_I = na_n$. Since $\wp(\alpha_m) = p'$, and $\lambda(\partial(\alpha_m)) = 0$, $\lim_{n\to\infty} a_n = p'$. For the $m_I$ points that lie in $I_{m,n}$, let $p''$ be the probability that a cell $g$ in $I_{m,n}$ remains empty. Then from (9),

$$p'' \leqslant \left(1 - \frac{1}{m^2\#I_{m,n}}\right)^{m_I} = \left(1 - \frac{1}{m^2}\frac{n^{d\delta}}{\#I_{m,n}}\frac{1}{n^{d\delta}}\right)^{na_n}. \tag{10}$$

As in the previous section, the limit of the expression on the right hand side of the above equals zero only when $\delta < 1/d$. As a result, the expected proportion of empty cells in the interior of $\alpha_m$ goes to zero. As the proportion is a non-negative bounded random variable, it itself goes to zero in probability. Since $\lambda(\partial(\alpha_m)) = 0$, this implies $\lambda(H^c_{m,n} \cap \alpha_m) \to 0$ as $n \to \propto$. Therefore, $\lambda(H^c_n \cap \alpha_m) \to 0$ in probability.

For any given $\varepsilon > 0$ and $0 < t < 1$, we can choose $M$ and $N$ such that whenever $m \geqslant M$, $n \geqslant N$ and $\delta < 1/d$, the probability

$$\begin{aligned} P(|\lambda(H^c_n \cap \alpha)| < \varepsilon) &\geqslant P(|\lambda(H^c_n \cap \alpha_m)| < \varepsilon/2) \\ &\geqslant P\left(\left|\lambda\left(H^c_{m,n} \cap \alpha_m\right)\right| < \varepsilon/2\right) \\ &\geqslant 1 - t. \end{aligned} \tag{11}$$

As (11) is true for arbitrary $\varepsilon$ and $t$, $\lambda(H_n^c \cap \alpha) \to 0$, in probability. Also, $\lambda(H_n \cap \alpha^c) \leqslant \lambda(B_n) \to 0$ as $n \to \propto$. Combining these results we have, $\lambda(H_n \Delta \alpha) \to 0$ in probability.

Thus, the following theorem is established.

**Theorem 4.** *Let $\Upsilon_n = \{x_1, \dots, x_n\}$ be i.i.d. observations from any continuous distribution $\wp$ having support $\alpha$ on which its density $f$ is positive. Then the s-shape $H(s_n)$ is a consistent estimator of $\alpha$ in d-dimensions under the conditions of Theorem* 3.

Hereafter, we assume that the distribution of the sample points is uniform. The minor difficulties that arise for other distributions can be covered by a more elaborate treatment.

### *3.3. Set consistency of the s-shape derivatives*

#### *3.3.1. Set consistency of $\overline{H}(s_n)$*

The smoothed induced hull $\overline{H}(s_n)$ in general, is a better representation of the shape of a dot pattern than s-shape. The consistency of $\overline{H}_n$ (which is an abbreviation of $\overline{H}(s_n)$) is analyzed. It can be easily verified that

$$\lambda(H_n) \leqslant \lambda(\overline{H}_n) \leqslant 3^d \lambda(H_n). \tag{12}$$

$$\text{As } \overline{H}_n \text{ is a superset of } H(s_n),\ \lambda(\overline{H}_n^c \cap \alpha) \to 0 \quad \text{in probability.} \tag{13}$$

The boundary error may increase in case of $\overline{H}(s_n)$. But as $\lambda(\overline{H}_n \cap \alpha^c) < 3^d \lambda(H_n \cap \alpha^c)$,

$$\lambda(\overline{H}_n \cap \alpha^c) \to 0 \quad \text{in probability.} \tag{14}$$

The above two equations result in the following theorem.

**Theorem 5.** *The smooth induced hull $\overline{H}(s_n)$ is a consistent estimator of $\alpha$ under the same condition imposed on Theorem* 3.

#### *3.3.2. Set consistency of elements of the s-shape spectrum*

Here we analyse the consistency of $H(s_n^t)$ for finite $t$'s. It is sufficient to establish that the expected proportion of empty cells completely in the interior of $\alpha$ goes to zero for sufficiently large $n$. The rest of the proof is similar to that of Theorem 3.

Let $H(s_n^t)$ be denoted by $H_n^t$ and $\#I_n^t$ be the number of cells completely in the interior of $\alpha$ with cell size $s_n^t$. Similar notations are also used for other related terms. Now,

$$s_n^{t+1} = n^{-\delta} \sqrt[d]{\lambda\left(H_n^0\right)} \prod_{i=1}^{t} \sqrt[d]{\frac{\lambda\left(H_n^i\right)}{\lambda\left(H_n^{i-1}\right)}}, \quad t \geqslant 1, \qquad \text{where } \lambda\left(H_n^0\right) = V_n. \tag{15}$$

We want to show that $H_n^{t+1}$ is a consistent estimator of $\alpha$ for any finite positive integer $t$.

With the initial cell length $s_n^1$, let the optimal window $W_n$ covering $\Upsilon_n$ be partitioned into $\#G_n^1(\approx n^{d\delta})$ cells and the consistency of $H_n^1$, i.e., $H(s_n)$ has already been established. For successive $t$, $\#G_n^t$ and $\#H_n^t$ will be similarly defined. Now consider, the case for $t = 1$

$$s_n^2 = n^{-\delta}\sqrt[d]{\lambda(H_n^1)} = n^{-\delta}\sqrt[d]{V_n}\sqrt[d]{\frac{\lambda(H_n^1)}{V_n}}.$$

By the consistency of the s-shape generated with cell length $s_n^1$,

$$\sqrt[d]{\frac{\lambda(H_n^1)}{V_n}} \to \sqrt[d]{p}. \tag{16}$$

And

$$\#G_n^2 \approx \frac{V_n}{n^{-d\delta}\lambda(H_n^1)} \approx \frac{n^{d\delta}}{p} \quad \text{for large } n. \tag{17}$$

Also

$$\lim_{n\to\infty}\frac{\#I_n^2}{\#G_n^2} = p \text{ implies}_{n\to\infty}\frac{\#I_n^2}{n^{d\delta}} = 1. \tag{18}$$

Thus, the expected proportion of empty cells among the $\#I_n^2$ in the interior of $\alpha$ is

$$\left(1 - \frac{1}{\#I_n^2}\right)^{na_n} = \left(1 - \frac{n^{d\delta}}{\#I_n^2}\frac{1}{n^{d\delta}}\right)^{na_n} \text{ where } \lim_{n\to\infty}a_n = 1. \tag{19}$$

For arbitrary large $n$, the above relation tends to $e^{-1}$ if $\delta = 1/d$; equals to 0 if $\delta < 1/d$; and equals to 1 if $1/d < \delta < 1$.

Thus, for $t = 1$, $H_n^{t+1}$ is a consistent set estimator under the conditions of Theorem 3. Note that for $t = 2$,

$$\#G_n^3 \approx n^{d\delta}\left[\frac{V_n}{\lambda(H_n^1)}\right] \times \left[\frac{\lambda(H_n^1)}{\lambda(H_n^2)}\right]. \tag{20}$$

Now, $V_n/\lambda(H_n^1)$ and $\lambda(H_n^1)/\lambda(H_n^2)$ tend to $1/p$ and 1, respectively, for large $n$. Thus, for large $n$

$$\#G_n^3 \approx \#G_n^2 \approx \frac{n^{d\delta}}{p}. \tag{21}$$

Subsequently, for $t = 2$, $H_n^{t+1}$ is also a consistent set estimator under the same conditions. Thus, by induction, the following theorem is established.

**Theorem 6.** *For any finite sequence in t, each element of the s-shape spectrum $\langle H(s_n^t)\rangle$ is a consistent estimator of $\alpha$ under the same conditions imposed on Theorem* 3.

### *3.3.3. Consistency of the clopen version of the s-shape $\overset{\circ}{\overline{H}}(s_n)$*

Let $\mathcal{A}$ denote the event that any non-zero position in the smooth binary projection $\overline{\mathcal{H}}(s_n) \subset \mathfrak{I}^d$, whose corresponding cell lies in the interior of $\alpha$, remains non-zero after opening. Now, to show the consistency of $\overset{\circ}{\overline{H}}(s_n)$, it is sufficient to establish that the probability of $\mathcal{A}$, $P(\mathcal{A}) \to 1$.

Consider a window, $w$ of size $5 \times 5 \times \ldots \times 5$ ($d$-tuple). Let $b$ be a non-zero (foreground) grid of $\overline{H}(s_n)$. It can be verified that to convert the value of $b$ to zero (background) by opening with the structuring element $\tau^d$, at least two zero valued positions should exist in the $w$ centered at $b$. (The number of required grids in background of $w$ increases with the dimensionality). Let $\eta =$ number of grids in the background lying within $w$ and adjacent to the foreground position, $b$ in $\overline{H}(s_n)$. For $i = 1, 2, \ldots, 5^d$, let $\chi_i$ be the characteristic function defined by

$$\chi_i = \begin{cases} 1 & \text{if } i\text{th grid is empty,} \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

represent the status of the $i$th cell in the $5 \times 5 \times \cdots \times 5$ ($d$-tuple) surrounding $b$.

By Markov's inequality,

$$P(\mathcal{A}) \leqslant P(\eta \geqslant 2) \leqslant \frac{E(\eta)}{2} = 5^d \frac{E(\chi_i)}{2} \to 0. \tag{23}$$

Thus, we get the following theorem.

**Theorem 7.** *The clopen version of the s-shape,* $\overset{\circ}{\overline{H}}(s_n)$ *is also a consistent estimator of* $\alpha$ *under the same condition imposed on Theorem* 3.

One interesting observation from (21) is that for large $n$ there is no significant change of the s-shapes in the spectrum for $t \geqslant 2$. Thus, whenever there is no apriori information on the support $\alpha$ (such as shape number, volume etc.), $H(s_n^2)$, which is the element of the s-shape spectrum after the second iteration is taken. For smoother rendering $\overline{H}(s_n^2)$ may be taken as the final output. However, if it is suspected that the data is noise-prone, then the clopen version $\overset{\circ}{\overline{H}}(s_n^2)$ should be adopted.

### 3.4. Error rate

It is crucial to have an idea of the order of error (in terms of the symmetric difference of volumes between $\alpha$ and $\alpha_n^*$) when the procedure is terminated at a particular value of $n$ and the corresponding estimate $\alpha_n^*$ 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 $\alpha$ separately.

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

$$E_I = \lambda(I_n) \times \left(1 - c_1 \frac{1}{n^{d\delta}}\right)^{c_2 n} \tag{24}$$

where $c_1$ and $c_2$ are positive constants.

The logarithm of the RHS of (24) is taken. Expanding $\log(1 - c_1(1/n^{d\delta}))^{c_2 n}$ and then exponentiating back, the leading term of $E_I$ is found to be $\exp\{-(c_1 \times c_2) n^{(1-d\delta)}\}$.

Let $\zeta(\partial(\alpha))$ denotes the $d$-dimensional surface area of $\alpha$ (in $d - 1$ dimensions). Then the error $E_B$ in the boundary satisfies

$$
\begin{aligned}
E_B &\leqslant \#B_n \times n^{-d\delta} \\
&\leqslant \left(\frac{\zeta(\partial(\alpha))}{n^{-(d-1)\delta}}\right) \times n^{-d\delta} \\
&\leqslant c_3 n^{-\delta}, \quad \text{where, } c_3 \text{ is a positive constant.}
\end{aligned}
$$

Note that the error in the boundary dominates that in the interior. Thus, the error in estimation is at most of order $\mathrm{O}(n^{-\delta})$. One important point to note is that the error is independent of the dimensionality.

## 4. Implementation

In the following, a simple algorithm of linear order time complexity is presented on s-shape, which can be applied for data in any finite dimensions. The method is somewhat similar to the clustering analysis method by binary morphology proposed by Postaire et al. [9], which may be consulted for detail computational analysis. The smooth versions of s-shapes are derived by basic morphological operators whose serial compositional properties are exploited, i.e., unit-dimensional arrays of size three are used as the structuring element for 2D as well as 3D data.

### 4.1. Computation of the s-shape

*Input*: Consider a set of $n$ $d$-dimensional (random) observations $\Upsilon_n = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ where $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,j}, \ldots, x_{i,d}]^{\mathrm{T}}$, $x_{i,j} \in \Re^1$ for $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, d$. The input is provided by a 2D array $x[n][d]$. The resolution parameter $\delta$ is also given.

Step 1. Find

$$O' \leftarrow \left\lfloor \min_i x_{i,1}, \min_i x_{i,2}, \ldots, \min_i x_{i,j}, \ldots, \min_i x_{i,d} \right\rfloor^{\mathrm{T}},$$

$$O'' \leftarrow \left\lfloor \max_i x_{i,1}, \max_i x_{i,2}, \ldots, \max_i x_{i,j}, \ldots, \max_i x_{i,d} \right\rfloor^{\mathrm{T}},$$

$$l \leftarrow [l_1, l_2, \ldots, l_j, \ldots, l_d]^{\mathrm{T}} \quad \text{where } l_j = \max_i x_{i,j} - \min_i x_{i,j} \text{ and } \lambda(W_n) \leftarrow \prod_{j=1}^{d} l_j.$$

Note that $W_n$ is the optimal isothetic hyper-rectangle covering $\Upsilon_n$ and its two opposite-diagonal corners are $O'$ and $O''$.

Step 2. Set $t \leftarrow 1$, $s_0 \leftarrow \lambda(H(s_0)) \leftarrow \lambda(W_n), \#H_t \leftarrow 0$.

Step 3. Find the initial side-length of cells of the s-shape $H(s_1)$: $s_1 \leftarrow n^{-\delta}(\sqrt[d]{\lambda(W_n)})$.

Step 4. Find $L \leftarrow [\lceil \frac{l_1}{s_t} \rceil, \lceil \frac{l_2}{s_t} \rceil, \ldots, \lceil \frac{l_j}{s_t} \rceil, \ldots, \lceil \frac{l_d}{s_t} \rceil]^{\mathrm{T}}$; $\#L \leftarrow \prod_{j=1}^{d} L_j$ where $L_j = \lceil \frac{l_j}{s_t} \rceil$.

Step 5. Initialize one 2D integer array $b[n][d]$, one single-dimensional integer array tag$[n]$ and one single-dimensional binary array $z[\#L]$ with zeros.

Step 6. For each $k = 1, 2, \ldots, n$, find

$$b_k = \left[ \left\lfloor \frac{\bar{x}_{k,1}}{s_t} \right\rfloor, \left\lfloor \frac{\bar{x}_{k,2}}{s_t} \right\rfloor, \ldots, \left\lfloor \frac{\bar{x}_{k,j}}{s_t} \right\rfloor, \ldots, \left\lfloor \frac{\bar{x}_{k,d}}{s_t} \right\rfloor \right]^{\mathrm{T}}$$

and

$$\text{index} \leftarrow \sum_{j=1}^{d} \left( b_{k,j} \times \prod_{m=1}^{j-1} l_m \right) + 1, \quad \text{where } b_{k,l} = \left\lfloor \frac{\bar{x}_{k,l}}{s_t} \right\rfloor, \bar{x}_{k,j} \leftarrow x_{k,j} - \min_i x_{i,j}.$$

$\text{tag}_k \leftarrow \text{index};$

If $(z_{\text{index}}$ is zero$)\{z_{\text{index}} \leftarrow 1;\ \#\mathcal{H}_t \leftarrow \#\mathcal{H}_t + 1.\}$

Step 7. Find the volume of the s-shape $H(s_t) : \lambda(H(s_t)) \leftarrow \#\mathcal{H}_t \times (s_t)^d$

Step 8. If

$$\left( \frac{|\lambda(H(s_{t-1})) - \lambda(H(s_t))|}{\lambda(H(s_{t-1}))} > \varepsilon \right) \wedge (t \leqslant Maxit)$$

$$s_{t+1} \leftarrow n^{-\delta}\left( \sqrt[d]{\lambda(H(s_t))} \right) \quad t \leftarrow t+1 \text{ and go to Step 4.}$$

*Output*: $b$, tag, $s_t, O'$ and $L(L_1, L_2, \ldots, L_j, \ldots, L_d)$.

The constant *Maxit* is two in the case of unsupervised learning. For details, see Section 3.3.3.

The information stored in $b$, tag, $s_t, O'$ and $L$ are sufficient for the s-shape $H(s_t)$ generation. The components of vector $L$ represent the discrete span of cells along axial directions. Rows in $b$ represents the location of non-empty grids of $\mathcal{H}(s_t)$, which is the lattice representation of $H(s_t)$. If $d$-dimensional unit hyper-cubes are placed at such locations and stretched by a scale factor of $s_t$, followed by a translation with $OO'$ ($O$ is the center of origin) then the result is the s-shape of the point set $\Upsilon_n$. The *tag* is used to find the mapping between cells of s-shape and the observations $\Upsilon_n$ by storing the serial index of each non-zero grid of $\mathcal{H}(s_t)$.

### 4.2. Smoothing via morphological transforms

Here we briefly describe the implementation of morphological transforms. As described in Section 2.1, only discrete basic morphological transforms (dilation and erosion) are applied and only $\tau^d$, the $3 \times 3 \times \cdots \times 3$ ($d$-tuple) array having all entries equal to 1 is required as the $d$-dimensional structuring element. However, $\tau^d$ is not directly applied. The serial composition properties of erosion and dilation are exploited for easy implementation of higher dimensional morphological transforms. Same conventional erosion and dilation results are achieved by applying directional respective transforms using only 1D structuring elements. For example, in case of 2D, dilation of a discrete set by $\tau^2$ is achieved by successive dilation with $\tau^1_{x_1}$ along the $x_1$ direction and $\tau^1_{x_2}$ along the $x_2$ direction. In Fig. 2, the directional dilation and the directional erosion are illustrated. When some observations are from outside $\alpha$ (noisy data) we apply clopen transform as the filter. Otherwise, closing is only applied on the s-shape for smoothing. It can be easily shown that $\mathcal{A}\bullet\mathcal{B} = \mathcal{A} \cup \mathcal{A}'$ where $\mathcal{A}' = \{y \in (\mathcal{A} \oplus \mathcal{B})/\mathcal{A}$ such that $\mathcal{B}_y \subseteq \mathcal{A} \oplus \mathcal{B}\}$ is a type of *conditional erosion* of $\mathcal{A} \oplus \mathcal{B}$ by $\mathcal{B}$ [10]. Thus, we can get the closed version of $\mathcal{A}$ by $\mathcal{B}$ just by adding those $y$ to $\mathcal{A}$ which are exclusively in $\mathcal{A} \oplus \mathcal{B}$ but not in original $\mathcal{A}$ such that $\mathcal{B}_y \subseteq \mathcal{A} \oplus \mathcal{B}$.

Two single-dimensional binary arrays $z'$ and $z''$ of size $\#L$ are used for morphological transforms. After computation of the s-shape, $z'$ is derived directly from *tag* that stores the serial indices of all non-zero grids of $\mathcal{H}(s_t)$. $z'_i$ equals one(zero) means the $i$th grid in $\mathcal{H}(s_t)$ is one (zero). In general, for any grid $(b_1, \ldots, b_d)$ in $\mathcal{H}(s_t)$, its serial index $i$ is $1 + \sum_{j=1}^{d} b_j \prod_{k=1}^{j-1} L_k$ ($L_k$ is the $k$th component of $L$). On the other hand, $z'_i$ corresponds to a grid in $\mathcal{H}(s_t)$ at the location $[\lfloor f_1 \rfloor, \ldots, \lfloor f_j \rfloor, \ldots, \lfloor f_d \rfloor]^{\mathrm{T}}$ where $f_d = (i-1)/u_{d-1},\ f_j = (\ldots((i - 1\%)u_{d-1})\%u_{d-2}\ldots)\%u_{d-(j+1)}/u_{j-1},\ u_k = \prod_{j=1}^{k} L_j$ for $j = 1, \ldots, d-1$, and $u_0 = 1$.

During first directional dilation of the original s-shape with $\tau^1_{x_1}$, only first coordinates of non-zero grids have to be considered. $z''$ is initialized by zeros. Without loss of generality, suppose $z'_i$ identifies a positive grid $(b_1, \ldots, b_d)$. Also, let $i'$ and $i''$ be the indices of $(b_1 - 1, \ldots, b_d)$ and $(b_1 + 1, \ldots, b_d)$ respectively. If $z'_{i'}$ equals zero then $z''_{i'}$ is set to one. Same criterion is applied for $i''$. $z' \leftarrow z' \vee z''$ gives the result of this directional dilation by $\tau^1_{x_1}$.

Only second coordinates of non-zero grids have to be considered for dilation by $\tau^1_{x_2}$. It is found similarly as in case of $\tau^1_{x_1}$. Thereafter, successive dilation by $\tau^1_{x_3}$, …, and finally by $\tau^1_{x_d}$ completes the process of dilation. The result is stored in $z'$.

Note that at this stage, $z''$ identifies the exclusive locations of dilated version of $\mathcal{H}(s_t)$ and corresponding grids are considered for conditional erosion to get the closed version. Suppose $(b_1, \ldots, b_d)$ with index $i$ is such a grid. Under (conditional) directional erosion along first dimension, if for the indices of $(b_1 - 1, \ldots, b_d)$ and $(b_1 + 1, \ldots, b_d)$ say $i'$ and $i''$, $z'_{i'} \wedge z'_{i''}$ is zero then both $z'_i$ and $z''_i$ are set to zeros. Like-wise, next (conditional) directional erosion along second direction (dimension) and rests of the (conditional) directional erosion along other dimensions are performed successively. Finally, the closed version is stored in $z'$.

Likewise, clopen transform is also performed with these two arrays.

Note that the serial compositions of erosion or dilation require less Boolean operations and are considerably faster than direct transforms. In 3D case, 27 Boolean operations are required on the whole data space, which can be accomplished by only a sequence of nine 1D erosions [9]. More generally, due to serial decomposition property, a $d$-dimensional morphological erosion or dilation is reduced to a cascade of $3 \times d$ 1D elementary filters.

## 5. Experimental results: data visualization

To demonstrate the effectiveness of our proposed techniques both as consistent set estimators and as shape descriptors, we experiment *s*-shape for several data examples. 2D digital images are studied to compare the true and estimated sets in the con-

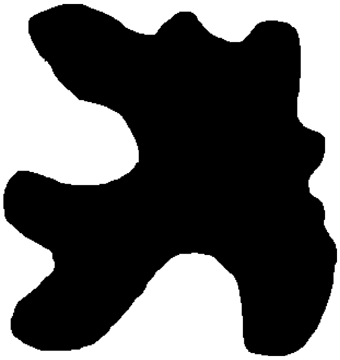

Fig. 1. A 'fish shaped' region.

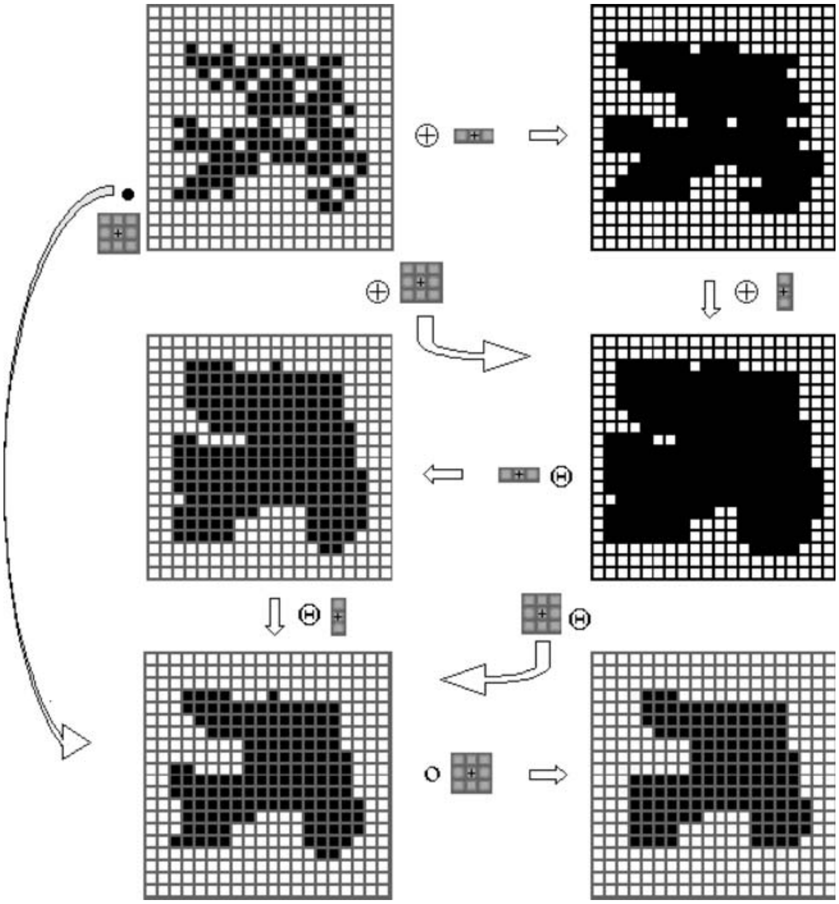

Fig. 2. Serial morphological compositional transformations used for smooth s-shape computations. Closing is the default smooth version of the object unless the data is noisy. Otherwise, clopen transform is used.

text of computational efficiency and to investigate the effect of additive noise. Examples with 3D images demonstrate the volume visualization from sample data. In the first set of experiments, the foreground in a (binary) digital image is considered as $\alpha$ and the area $\lambda(\alpha)$ is measured by its total number of object pixels. On the other hand, in 3D, models are generated from basic parametric shapes such as parallelepiped, sphere and cylinder. These models are treated as $\alpha$. Volume of $\alpha$, $\lambda(\alpha)$ is calculated from basic constituent parametric shapes.

In Fig. 1, $\lambda(\alpha)$ is 63,903. Random sample points are taken as shown in Figs. 3A–C with $n = 100$ ($\approx$0.15% data), 1500 ($\approx$2.34% data) and 3000 ($\approx$4.69% data), respectively. For $\delta = 0.45$, the estimators $H(s_n)$ and $\overline{H}(s_n)$ are presented in Figs. 4A–C and 4D–F, respectively. Results for $\delta = 0.49$ are shown in Fig. 5. The ratio of $\lambda(\alpha_n^*\Delta\alpha)$ to $\lambda(\alpha)$ are plotted against the sample size for $\delta = 0.45$ and 0.49 in Figs. 6A and B, respectively. The asymptotic convergence of $\alpha_n^*$ is readily understood despite the limitation due to finite quantization. In terms of number of pixels, less than 7% of the total data is sufficient for the convergence. As far as set estimation is concerned, the smoothing leads to a substantial improvement for the case $\delta = 0.49$, but not for $\delta = 0.45$.

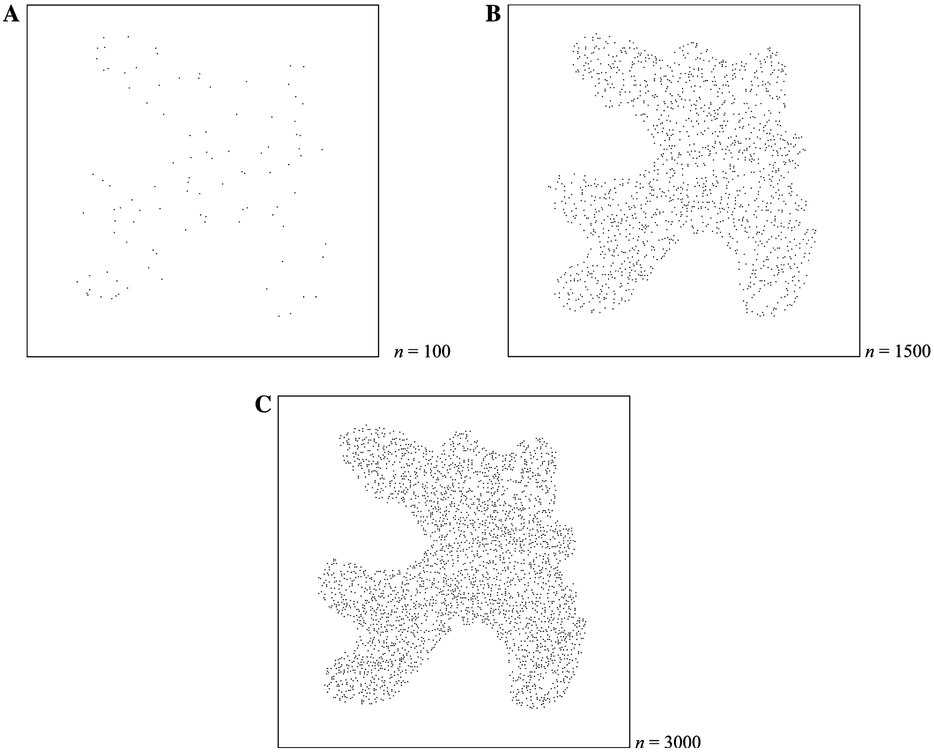

Fig. 3. Random samples from $\alpha$ of Fig. 1.

A major competitor of our set estimator is based on the MST. It is the only other set estimator which satisfies scale equivariance property and remains consistent when observations are drawn under any continuous distribution and might be used as a shape descriptor subject to number of components in the region of support being known and the data being noise-free. For s-shape based set estimator no such con-

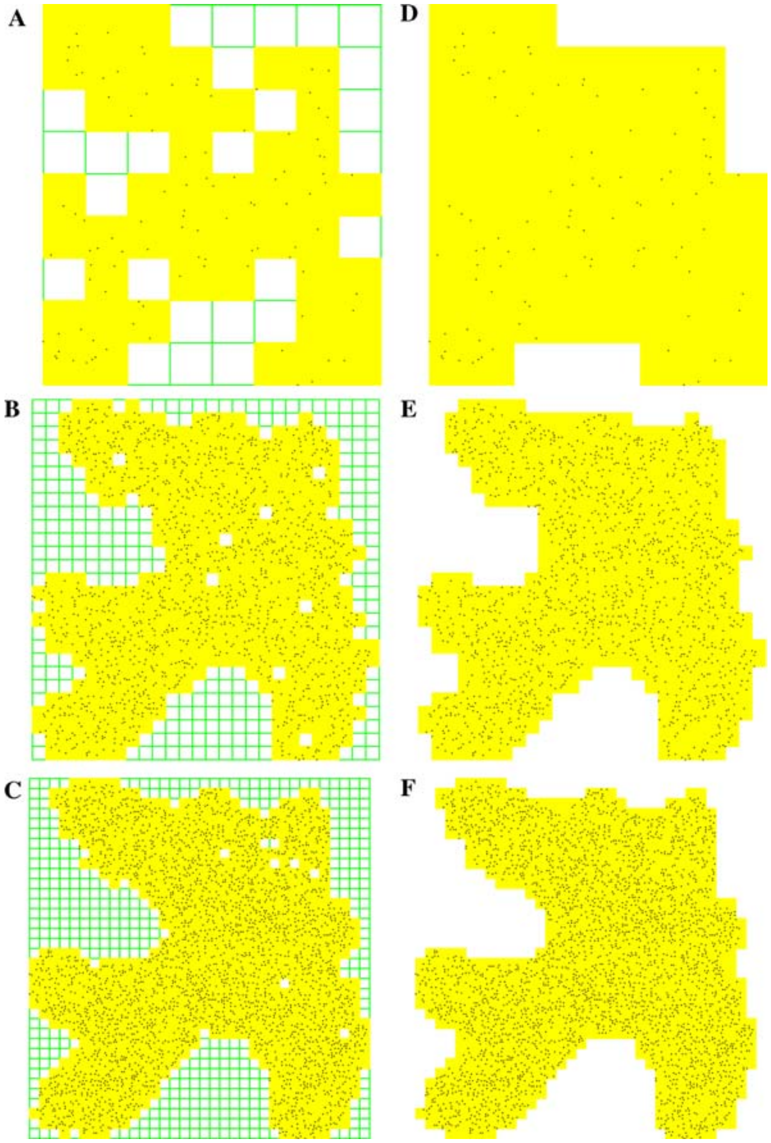

Fig. 4. The asymptotic convergence for $\delta = 0.45$.

straints are necessary. In Fig. 7, a run-time comparison is performed on basic modules (without considering the display module) of both the estimators where the region of support $\alpha$ is as in Fig. 1. In this example, as expected, s-shape computation is considerably faster than MST.

Fig. 8 presents two more examples where s-shape is applied as shape descriptor. The hole (Fig. 8A) and disconnected components (Fig. 8D) are correctly recovered.

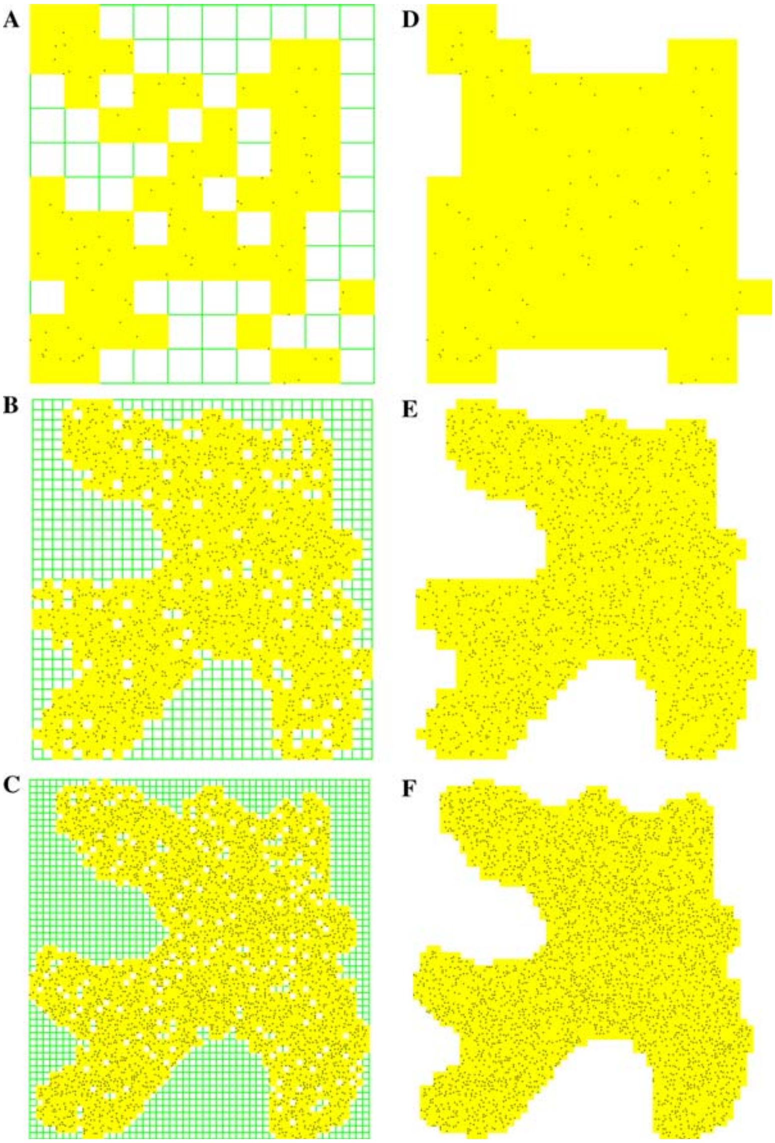

Fig. 5. The asymptotic convergence for $\delta = 0.49$.

Sample sizes of 300 and 1500 are taken in both sets which, in terms of number of pixels, only about 0.6 and 3% for the first set and 1.12 and 5.60% for the other set, respectively. All output figures represent smoothed s-shapes due to closing with $\delta$ fixed at 0.49. All these results also indicate that s-shape might be utilized for data compression.

The robustness of the s-shape based class of set estimators is demonstrated in Fig. 9. The input patterns are noisy [25]. In all these cases signal ($\alpha$) to random noise ratio, SNR, are fixed to 10 db and the used estimator is $\overline{H}(s_n^2)$ (with $\delta$ fixed at 0.49).

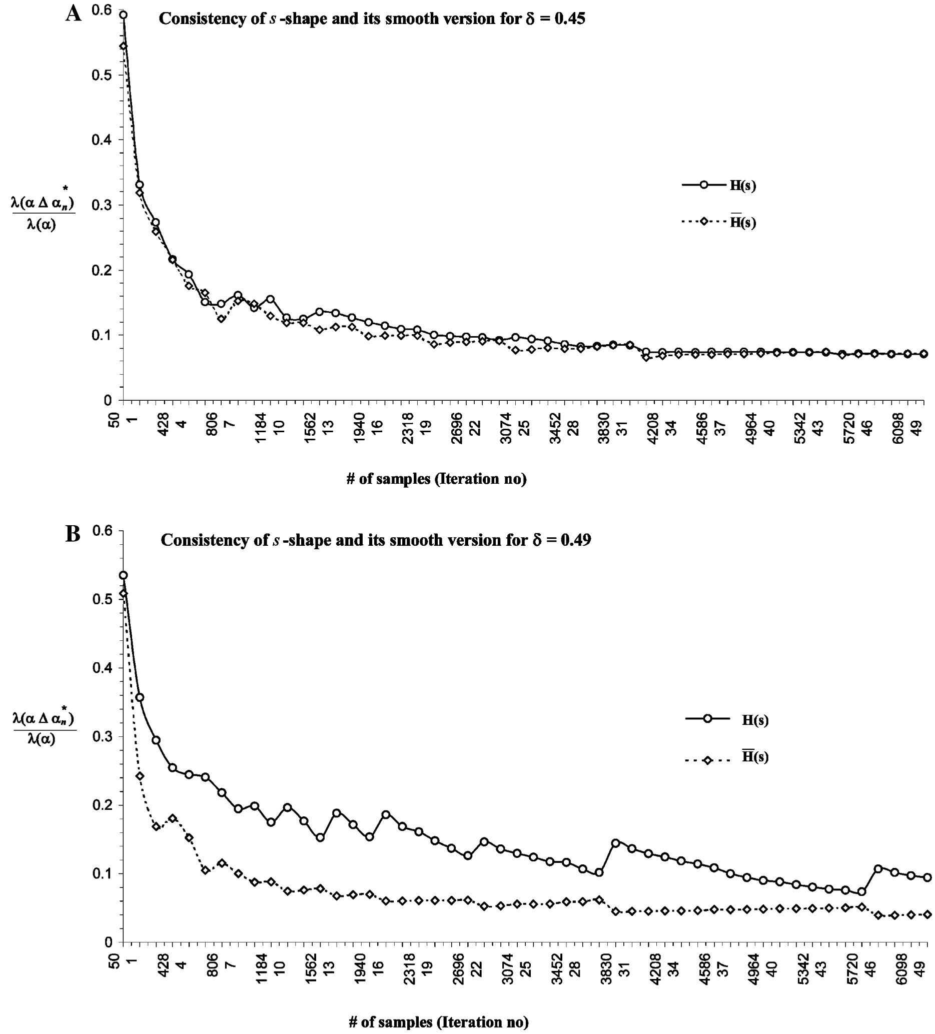

Fig. 6. Plots showing asymptotic convergence of proposed estimator in 2D.

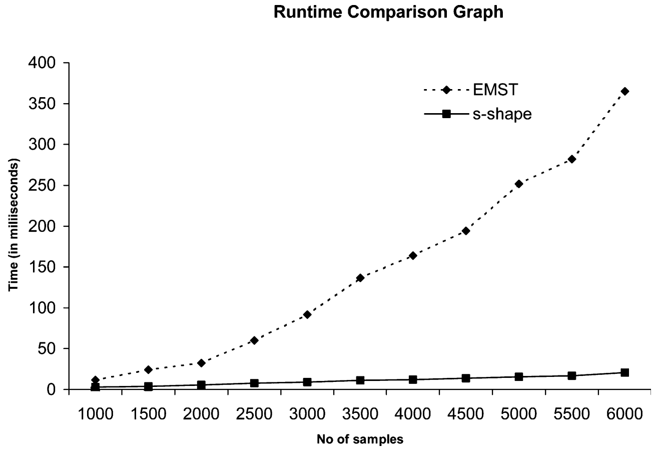

Fig. 7. A run-time comparison between basic modules of MST based consistent set estimation and that of s-shape. Random samples are drawn from the data of Fig. 1.

### 5.1. Role of $\delta$

It is clear that the choice of $\delta$ has considerable impact on the resulting s-shape. Here we try to analyze its impact in 2D data. For smaller values of $\delta$, the boundaries of $\alpha_n^* = H(s_n)$ are cruder—so much so that the s-shapes for $\delta$ in the range $(0, 0.45)$ appear to be of little practical utility. For larger values of $\delta$, on the other hand, the figure exhibits larger number of inconsistent holes (compare Figs. 4 and 5). In the particular case $\delta = 0.5$, the proportion of the area formed by the union of these holes with respect to the area of the region under estimation converges to a fixed non-zero constant so that consistency fails to hold. This suggests that 'smoothing' may be more useful for s-shapes obtained with values of $\delta$ close to $1/d$ i.e., 0.5. For a given $n$, larger values of $\delta$ lead to small values of $\lambda(\alpha_n^* \cap \alpha^c)$ and smaller values of $\delta$ lead to small values of $\lambda((\alpha_n^*)^c \cap \alpha)$, i.e., values of $\delta$ near opposite ends of the allowable range are more efficient in reducing complementary components of the symmetric difference.

Note that larger values of $\delta$ reduce the dominant boundary error. On the whole, in general dimensional case, it appears that when single values of $\delta$ have to be recommended, it should be always close to $1/d$. We take it as $(1 - \sigma)1/d$ where $\sigma$ is sufficiently small and unless otherwise mentioned, in all our experiments $\sigma \approx 0.1$. When coupled with smoothing, including the case of noisy data, the value of $\delta$ is further increased by taking the value to $(1 - \sigma^2)1/d$.

### 5.2. 3D volume visualization by s-shape

Here we demonstrate how volume visualization can be achieved with s-shape. Each $\alpha$ is generated from some basic 3D shapes such as sphere, cylinder, parallelepiped, cone, etc. In each case, $n$ randomly selected points with real coordinates are

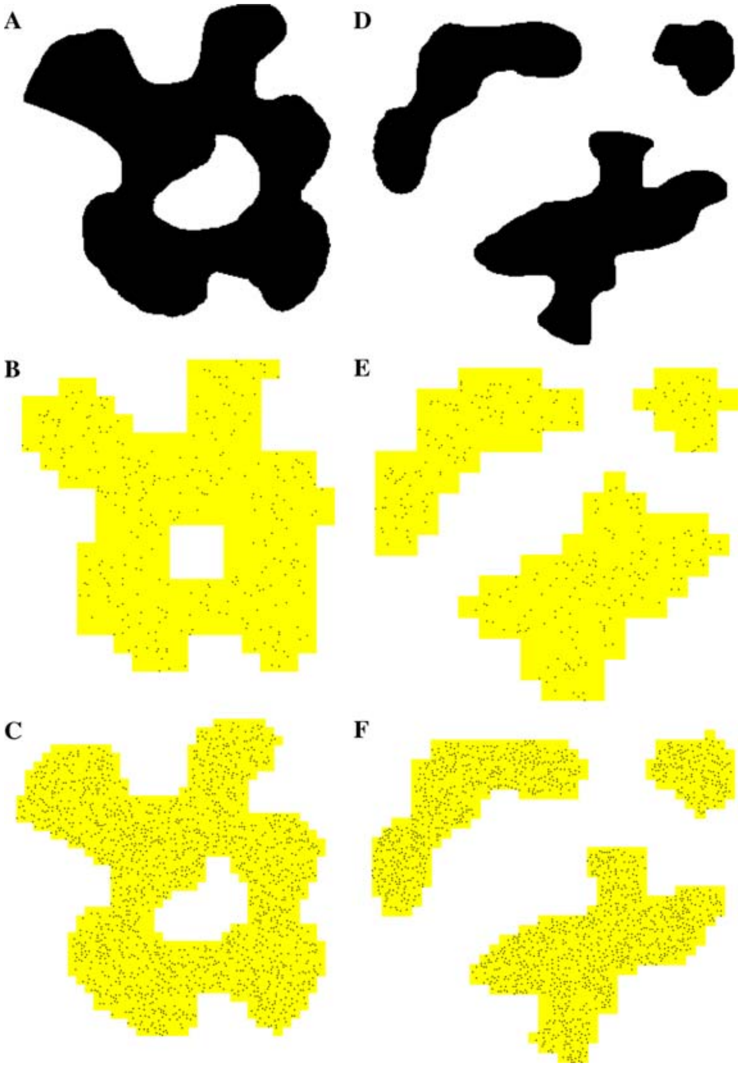

Fig. 8. Two more examples where smooth s-shape based set estimators extract pattern shapes.

sampled and their s-shapes are computed. The simulations are performed in a SUN workstation. Standard shading techniques available in MatLab6 are applied for depth perception. Fig. 10 shows three 3D models from where representative points are randomly drawn. In the left figure $\alpha$ is a torus, in the middle, it is a collection of models of some simple industrial tools and the model in the right one represents an extreme case where an inner sphere is completely hidden within a bigger hollow sphere. Representative points from these models, their original s-shapes, smooth versions and randomly taken slices are shown in Fig. 11. In Fig. 11C, though the density

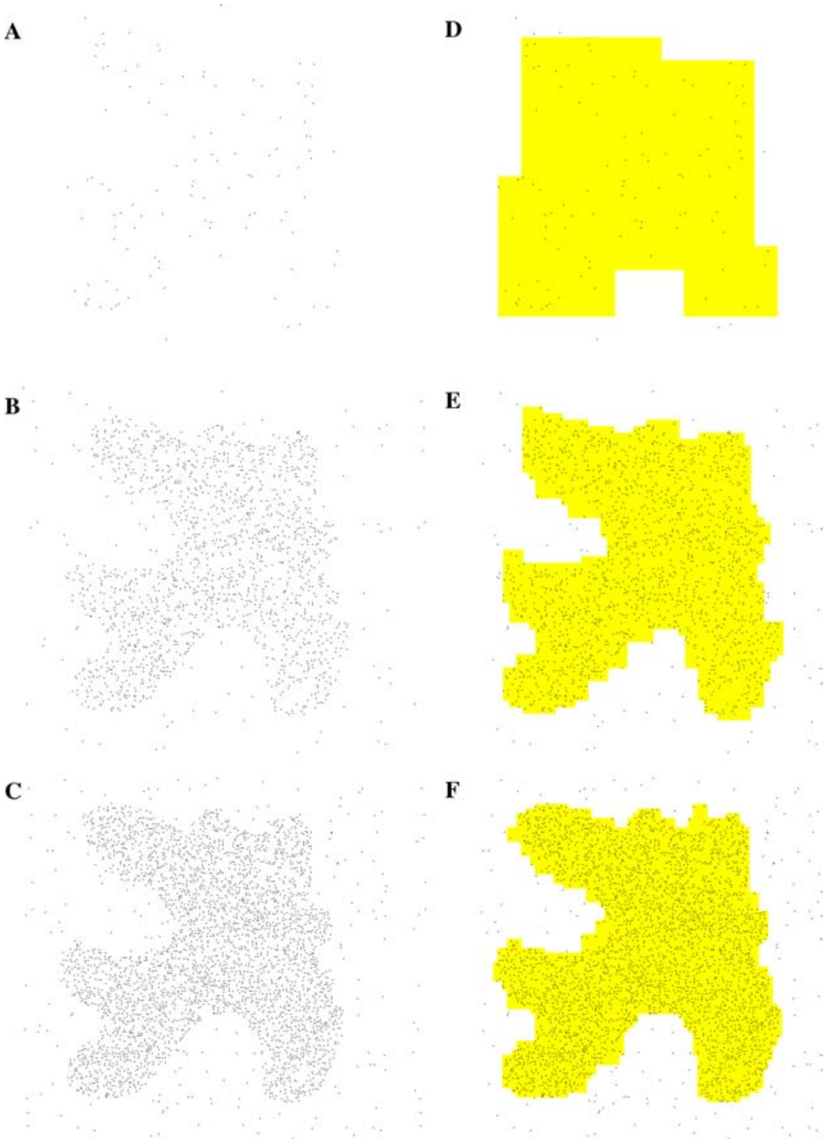

Fig. 9. Robustness of the class of proposed set estimators. (SNR = 10 db).

of points are equal for both the objects in the sample but inner part seems more denser due to overlapping of points from inner sphere and outer one.

It is apparent that finer rendering is achieved for larger $n$. However, at any stage, the shape is only a rough approximation, good enough to estimate the underlying zone and thus may be used as primitive solid modeling. The decisive advantage of s-shape based volume rendering is that the process is considerably fast due to linear

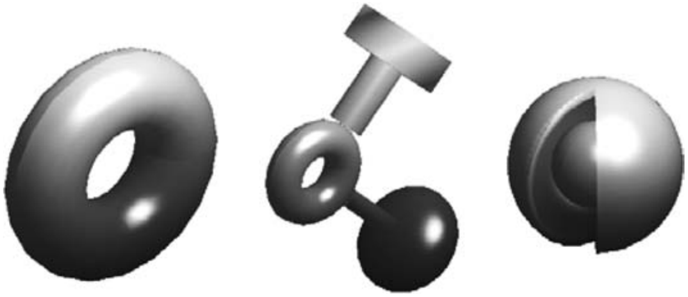

Fig. 10. Three models that demonstrate s-shape based volume visualization.

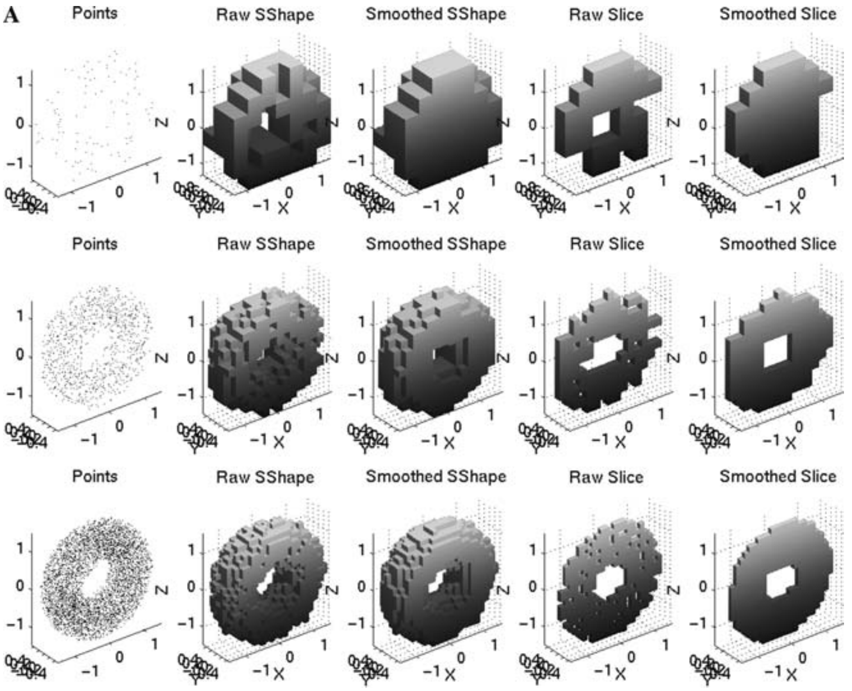

Fig. 11. (A–C) The above three figures show how s-shape can be used to visualize volume from finite observations. (In all these three sets $n = 100$, $n = 1000$, $n = 5000$). Note that slicing of the s-shape enables to see the 'inside' of the object. Particularly, in the last figure density of points are equal for both the objects but inner part seems more denser due to overlapping of points from inner sphere and outer one.

order computation. In Fig. 12, the run-time analysis for 3D point sets is performed. Here $\alpha$ is a unit cube. From the graph, it is apparent the s-shape based volume visualization is considerably fast. One important advantage of s-shape based volume visualization is that it allows looking inside the volume by taking horizontal and

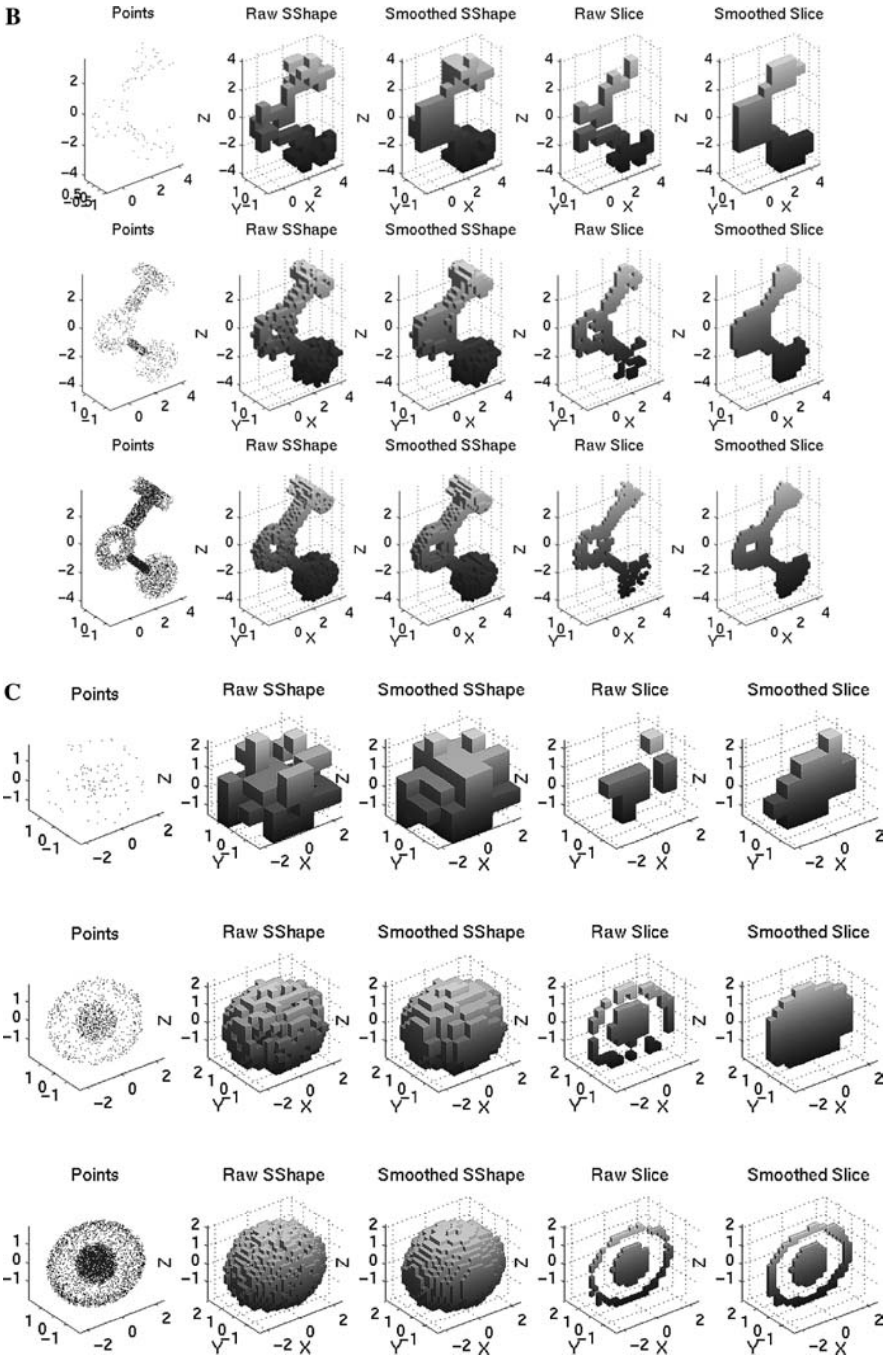

Fig. 11. (*continued*)

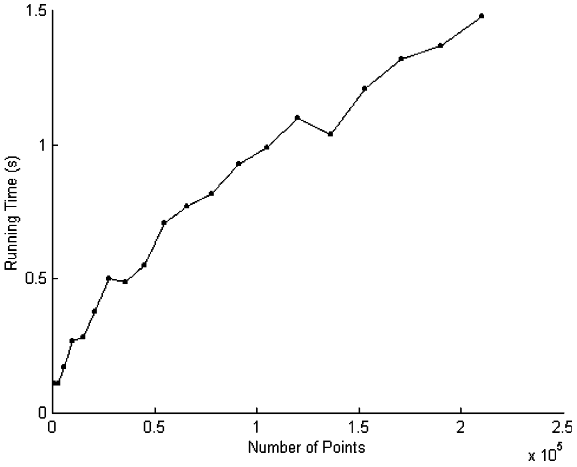

Fig. 12. Run-time plot for (smooth) s-shape generation ($\alpha$ is a unit cube).

vertical slices as shown in Fig. 11. This is not possible by most of the existing surface and volume reconstruction methods such as based on Voronoi/Delaunay tessellations.

## 6. Summary and discussion

A new set estimator called s-shape is presented for high dimensional data, which is applied for data visualization. It is a compact representation of equal cells of appropriate sizes that surround the sample. It is consistent under any continuous distribution in finite dimensions. The set consistency holds good even when there is no apriori information on whether the support set is single or multi-connected or the sample is corrupted by noise. These cannot be achieved by other exiting set estimators known to us including the one based on the MST.

Data exploration for finding the object of interest and their visualization from sample is an important area in various applications. For minute reconstruction, often million of data (points) are required and there exist methods that can minutely reconstruct the surface from sample but are computationally intensive [18–20]. However, if the object of interest is closely surrounded by other objects in the region of support (the area of exploration), it is very difficult to identify the object of interest despite sufficient amount of sampling. Also, looking inside of a single object is not possible by these methods. The s-shape based estimator can be used to complement these descriptors. Particularly, as demonstrated, it is useful as a basic volume visualizer. As it renders very fast and supports slicing, localization of the area of the object of interest within the region of support, even with a moderate-sized sample, can be efficiently managed. In that localized area, if additional representative points are accumulated and data scattered elsewhere are removed then this type of 'guided sampling' prevails overall data reduction. Finally, restored points in the border cells of the s-shape may be interpolated by a surface rendering method for reconstruction of the object in high resolution, such as [19].

In future, we will extend the present work in two directions. First, we will investigate the applicability of similar techniques for set estimation as well as shape reconstruction when the points are not in crisp states but have intensity within a certain finite range. Next, a complete graphical user interface will be also developed. It will support volume visualization, slicing to look within the data to identify the object of interest, guided sampling as discussed above for final smooth and detailed surface rendering.

## Acknowledgments

The first author acknowledges the Coordinated Science Laboratory and ECE Department of University of Illinois at Urbana Champaign for providing a Postdoctoral research fellowship for executing the experimental part of the presented work at Beckman Institute for Advanced Science and Technology. The authors are also thankful to Professor Narendra Ahuja, University of Illinois for his encouragement for pursing this work.

## References

[1] H. Edelsbrunner, E.P. Mücke, Three-dimensional alpha shapes, ACM Trans. Graphics 13 (1994) 43–72.
[2] P. Bajcsy, A. Ray Chaudhuri, Benefits of high resolution SAR for ATR of targets in proximity, Radar Conf. Proc. of the IEEE (2002) 29–34.
[3] A. Okabe, B. Boots, K. Sugihara, Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, Wiley, New York, 1992.
[4] J.C. Russ, Image Processing HandBook, CRC Press, Ann Arbar, 1995.
[5] U. Grenander, Abstract Inference, Wiley, New York, 1975.
[6] P.L. Combettes, Foundations of set theoretic estimation, Proc. IEEE 81 (2) (1993) 182–208.
[7] A. Ray Chaudhuri, B.B. Chaudhuri, S.K. Parui, A novel approach to computation of the shape of dot pattern and extraction of its perceptual border, Comput. Vision Image Understanding 68 (3) (1997) 257–275 (doi: 10.1006/cviu.1997.0550).
[8] A. Ray Chaudhuri, A. Basu, S. Bhandari, B.B. Chaudhuri, An efficient approach to consistent set estimation, Sankhya–B 61 (1999) 496–513.
[9] J.G. Postaire, R.D. Zhang, C.L. Botte, Cluster Analysis by binary morphology, IEEE Trans. Pattern Anal. Mach. Intell. 15 (2) (1993) 170–180.
[10] R. Haralick, L. Shapiro, Computer and Robot Vision, vol. 1, Chapter 5, Addison-Wesley, 1992.
[11] D.P. Mandal, C.A. Murthy, S.K. Pal, Determining the shape of a pattern class from sampled points in $\Re^2$, Int. J. Gen. Syst. 20 (1992) 307–339.
[12] 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.
[13] T.M. Apostal, Mathematical Analysis, Addision-Wesley, 1971.
[14] H. Edelsbrunner, D.G. Kirkpatrick, R. Seidel, On the shape of a set of points in the plane, IEEE Trans. Informat. Theory IT-29 (1983) 551–559.
[15] N. Ahuja, M. Tuceryan, Extraction of early perceptual structure in dot patterns: integrating region, boundary, and component Gestalt, Comput. Vision, Graphics, Image Process. 48 (1989) 304–356.

[16] M. Worring, A.W.M. Smeulders, Shape of an arbitrary finite point set in $\Re^2$, J. Math. Imaging Vision 4 (2) (1994) 151–170.

[17] D.R. Cox, D.V. Hinkley, Theoretical Statistics, Chapman & Hall, London, 1974.

[18] N. Amenta, M. Bern, M. Kamvysselis, A new Voronoi-based surface reconstruction algoritm, in: Proc. SIG-GRAPH '98 (1998).

[19] D. Attali, *r*-Regular shape reconstruction from unorganized points, in: 13th ACM Symp. on Computational GeometryissuenrJune (1997) 248–253.

[20] M. Melkemi, L. Chen, D. Vandorpe, Shapes of weighted sets of points, in: ICPR 2000, Barcelone, Espagne, September 2000.

[21] A. Ray Chaudhuri, A. Basu, S. Bhandari, B.B. Chaudhuri, Consistent set estimation in *k*-dimensions: an efficient approach, in: Lecture Notes in Computer Science Series, vol. 1451, Springer-Verlag, Berlin, 1998, pp. 667–676.

[22] K. Voss, Images, objects and surfaces in $Z^n$, Int. J. Pattern Recog. Artif. Intell. 5 (1991) 797–808.

[23] J. Serra, Image Analysis and Mathematical Morphology, Academic Press, New York, NY, 1982.

[24] R.M. Haralick, S.R. Sternberg, X. Zhuang, Image analysis using mathematical morphology, IEEE Trans. Pattern Anal. Mach. Intell. 9 (1987) 523–550.

[25] X. Zhou, R. Gordon, Generation of noise in binary images, CVGIP: Graphical Models and Image Processing 53 (5) (1991) 476–478.