A DATA DRIVEN PROCEDURE FOR DENSITY ESTIMATION WITH SOME APPLICATIONS

D. CHAUDHURI,* B. B. CHAUDHURI*† and C. A. MURTHY‡

* Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India

‡ Machine Intelligence Unit, Indian Statistical Institute, 203, B. T. Road, Calcutta 700 035, India

(*In revised form* 3 *July* 1995; *received for publication* 28 *February* 1996)

Abstract—This paper deals with the probability density estimation using a kernel-based approach where the window size of the kernel is found by a data-driven procedure. It is theoretically shown that, under certain assumptions, the estimated densities on bounded sets can be asymptotically unbiased when the width of window is obtained from the minimal spanning tree of the observed data. The theoretical development initially carried out on $\mathcal{R}^2$ is applicable to higher dimensional spaces. The results are experimentally verified on bounded sets with different types of distributions. The behaviour of the estimator in the case of the unbounded set as in that for Gaussian density is also experimentally seen to be good. Some applications of the proposed density estimation technique is demonstrated. One application is the representative point detection algorithm, which can be applied for data reduction and outlier rejection. Another application involves detection of border points of a dot pattern as well as finding a thinned version of the dot pattern.

Probability density estimation Kernel method Window selection
Minimal spanning tree Bounded set Asymptotically unbiased estimator
Representative point

1. INTRODUCTION

This paper deals with the non-parametric probability density estimation from a given set of data in multi-dimensional space. Density estimation is useful and important in statistical approaches to various problems of image processing, pattern recognition and artificial intelligence.

Perhaps the earliest attempts on non-parametric density estimation were based on histograms with uniform and random partitions.[1] The method of splines, such as histosplines, is another approach to obtain a smooth density mapping.[2] Reports on these two techniques consider mostly one- and two-dimensional data and their applicability to problems in arbitrary dimension is limited. The method of kernels[3,4] and the k-nearest neighbours are two most popular estimation procedures that can be applied to data of arbitrary dimension. Among others, the method of orthogonal expansion using, say, trigonometric basis functions can be mentioned. Functions such as Hermite polynomials may also be used for the expansion.[7] The methods of delta sequences,[9] penalty functions[10] and stochastic approximation[11] are among the other techniques of density estimation. For a general review on the topic, see Wertz[12] and Prakasa Rao.[13]

The work presented here is related to the method of kernels. The idea behind the kernel method is as follows. A window function satisfying the conditions of probability density function is chosen. For a multivariate datum $\mathbf{x}$, the density estimate $f_n(\mathbf{x})$ is defined as the average of the window function values at $\mathbf{x}$ with origins centred at the data points, where n is the total number of points. A parameter h_n, which is the width of the window, decides how well the local variation in the actual density f will be reflected in the estimated density f_n. When the number of samples tend to infinity, f_n becomes an asymptotically unbiased and consistent estimator of f under certain restrictions on h_n.

Clearly, the choice of h_n has a major effect on $f_n(\mathbf{x})$. If h_n is too large, the estimate will suffer from too little resolution. If h_n is too small, the estimate will suffer from too much statistical variability. In this paper, the window parameter h_n is found by a data-driven procedure. It can be understood that the choice of h_n should depend on n as well as the relative distance between points in the data set. The dependency of n and interpoint distances is combined here by considering h_n as a function of the length of the Minimal Spanning Tree (MST) drawn on the data set. The choice of the window size h_n and the proposed density estimation procedure have been described in Section 2.

In the original kernel methods of Parzen[3] and Cacoullos[4] h_n is a sequence of numbers, while in our case h_n will be a random sequence. Hence, we should have $h_n \to 0$ and $nh_n^2 \to \infty$ in probability as $n \to \infty$.

Thus, the consistency and asymptotic unbiasedness of our estimation do not directly follow the arguments of Parzen and Cacoullos and should therefore be proved. The theoretical aspects of the proposed scheme are treated in Section 3. Experimental results on simulated data for specific densities are presented in Section 4. The applications of the density estimation method to representative points detection, border points detection and generation of a thinned version of the dot pattern are also presented in Section 5.

2. PROPOSED DENSITY ESTIMATION APPROACH

Consider the given set of points $D_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathscr{R}^q$. The kernel-based approach for density estimation is as follows.

Suppose $K(\mathbf{y})$ is a Borel scalar function on $\mathscr{R}^q$ such that:

$$\sup_{\mathbf{y}\in\mathscr{R}^q} |K(\mathbf{y})| < \infty \qquad \int_{\mathscr{R}^q} |K(\mathbf{y})|\,d\mathbf{y} < \infty$$
$$\lim_{|\mathbf{y}|\to\infty} |\mathbf{y}|^q K(\mathbf{y}) = 0 \qquad \int_{\mathscr{R}^q} K(\mathbf{y})\,d\mathbf{y} = 1, \tag{1}$$

where $|\mathbf{y}|$ denotes the length of the vector $\mathbf{y}$ on $\mathscr{R}^q$. $K(\mathbf{y})$ is termed the *kernel* of the density estimator. Let $f(\mathbf{y})$ be the actual density function on $\mathscr{R}^q$. Let:

$$f_n(\mathbf{x}) = \frac{1}{nh_n^q} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right), \tag{2}$$

where $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are independent and identically distributed random vectors following the density f and $\{h_n\}$ is a sequence of positive constants satisfying $h_n \to 0$ and $nh_n^q \to \infty$ as $n \to \infty$. Then at every continuity point $\mathbf{x}$ of f, we have:

$$\lim_{n\to\infty} E[f_n(\mathbf{x})] = f(\mathbf{x}).$$

Note that, if h_ns are sequence of numbers satisfying $h_n \to 0$ and $nh_n^q \to \infty$ as $n \to \infty$, then $f_n(\mathbf{x})$ is an asymptotically unbiased and consistent estimate of $f(\mathbf{x})$ for every q.[4]

We propose the following approach for density estimation where h_n is computed by a data-driven procedure. We want to combine the number of data and the interpoint distance through the MST introduced by the data.

Consider a set A which is path connected, compact, $cl(Int(A)) = A$ in $\mathscr{R}^q$ containing the set D_n of random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Construct a *minimal spanning tree* (MST)[5,6] on D_n where each $\mathbf{x}_i$ denotes a node and if $\mathbf{x}_i$ and $\mathbf{x}_j$ are connected by an edge then the edge weight is defined as the Euclidean distances between $\mathbf{x}_i$ and $\mathbf{x}_j$. Let the sum of the edge weights of MST of D_n be termed the length l_n of the MST. Then it can be shown (see Appendix) that $(l_n/n) \to 0$ in probability and $l_n \to \infty$ in probability as $n \to \infty$. Now, if we take:

$$h_n = \left(\frac{l_n}{n}\right)^{1/q}, \quad q \geq 2, \tag{3}$$

then $h_n \to 0$ in probability and $nh_n^q \to \infty$ in probability as $n \to \infty$. For any $q \geq 2$, we can obtain a sequence of random variables $h_1, h_2, \ldots$ on the basis of $\mathbf{x}_1, \mathbf{x}_2, \ldots$ such that $h_n \to 0$ in probability and $nh_n^q \to \infty$ in probability as $n \to \infty$. We take the kernel $K(\mathbf{x})$ as:

$$K(\mathbf{x}) = \begin{cases} \dfrac{1}{2^q} & \text{if } |x_j - x_{ij}| \leq 1 \quad \forall j = 1, 2, \ldots, q \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})'$ and $\mathbf{x} = (x_1, x_2, \ldots, x_q)'$, and $'$ denotes the transpose. Note that $K(\mathbf{x})$ satisfies the properties (1) of a kernel given above.

Using equations (2)–(4) the density at every $\mathbf{x} \in \mathscr{R}^q$ can be estimated.

It may be mentioned here that h_n as defined in equation (3) was also found to be useful in set estimation problems.[14]

3. THEORETICAL STUDY

This section deals with the theoretical properties of h_n. We shall define two sets $\mathscr{A}$ and $\mathscr{C}_A$ for the sake of mathematical clarity. At first, we concentrate on 2-D (two-dimensional) space. Here we rigorously prove that if we use equations (2)–(4) for density estimation, then the estimate will be consistent and asymptotically unbiased as stated by Theorem 1 and Theorem 2 below. However, to prove the theorems several lemmas are necessary. At first, we state the lemmas (with proofs given in Appendix). Next, the theorems are stated and proved.

3.1. *Analysis in 2-D space*

Let $\mathscr{A} = \{A \subseteq \mathscr{R}^2$: A path connected, compact, $cl(Int(A)) = A$, $Int(A)$ is path connected and $\lambda(A \cap cl(A^c)) = 0$, where λ denotes the Lebesgue measure in $\mathscr{R}^2\}$. $\mathscr{A}$ is the collection of all those subsets of $\mathscr{R}^2$ which support the density function. Let $\mathscr{C}_A = \{f: \mathscr{R}^2 \to [0, \infty)$: f is continuous on $Int(A)$, $f(\mathbf{x}) > 0 \ \forall \mathbf{x} \in Int(A)$, $\int_A f\,d\lambda = 1$, $\exists T > 0$ such that $|f(\mathbf{x})| \leq T \ \forall \mathbf{x} \in A$, where T is finite, $f(\mathbf{x}) = 0 \ \forall \mathbf{x} \in A^c\}$, $\forall A \in \mathscr{A}$, i.e. $\mathscr{C}_A$ is the set of all density functions with support A. The results concerning the consistency and asymptotic unbiasedness of $f_n(\mathbf{x})$ are stated below.

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \ldots$ be independent and identically distributed random vectors with density f, i.e. $\exists A \in \mathscr{A}$ such that $f \in C_A$. Let $D_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$. Let l_n be length of MST of D_n. Let $h_n = \sqrt{(l_n/n)}$ and

$$K(\mathbf{y}) = \begin{cases} 1/4 & \text{if } |y_1| \leq 1 \text{ and } |y_2| \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{y} = (y_1, y_2)'$.

Let the distribution function of h_n be represented by $\beta_n(\alpha)$, i.e.

$$P(h_n \leq \alpha) = \beta_n(\alpha) \quad \forall \alpha > 0.$$

Lemma 1. $P(nh_n^2 > M) \to 1$ as $n \to \infty$ for every $M > 0$.

Lemma 2. $P(h_n < \varepsilon_1) \to 1$ as $n \to \infty$ for every $\varepsilon_1 > 0$.

Lemma 3.

$$E[f_n(\mathbf{x})] = E\left[\frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_1}{h_n}\right)\right].$$

Proofs of Lemmas 1–3 are given in the Appendix.

Construction. We can construct two sequences of numbers t_n and γ_n such that for every $\varepsilon > 0$, $\exists M_0 > 0$ such that:

$$\text{(i) } P(t_n \geq h_n \geq \gamma_n) \geq 1-\varepsilon \quad \forall \quad n \geq M_0 \tag{5}$$
$$\text{(ii) } t_n \to 0 \text{ and } nt_n^2 \to \infty \quad \text{as } n \to \infty \tag{6}$$
$$\text{(iii) } \gamma_n \to 0 \text{ and } n\gamma_n^2 \to \infty \quad \text{as } n \to \infty \tag{7}$$
$$\text{(iv) } (t_n/\gamma_n) \to 1 \quad \text{as } n \to \infty \tag{8}$$

The validity of the construction is also proved in the Appendix.

Lemma 4.

$$\text{(i) } E\left[\frac{1}{nt_n^2}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{\gamma_n}\right)\right] \to f(\mathbf{x}) \text{ as } n \to \infty$$

$$\text{(ii) } E\left[\frac{1}{n\gamma_n^2}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{t_n}\right)\right] \to f(\mathbf{x}) \text{ as } n \to \infty$$

The proof of Lemma 4 is straightforward[6] under the condition (8).

Let us define $\mathcal{N}_x(\alpha) = \{\mathbf{y}: |x_1 - y_1| \leq \alpha,\ |x_2 - y_2| \leq \alpha,\ \mathbf{y} = (y_1, y_2)'\}$ where $\mathbf{x} = (x_1, x_2)'$.

Assumption 1. Let $\exists M_1 > 0$ such that $P(\mathbf{x}_1 \in \mathcal{N}_x(\alpha) / h_n = \alpha) \leq \alpha^2 M_1$ $\forall n \geq M_2 > 0$ and $\forall \alpha$.

Let us define S_n: $\mathcal{B}(\mathcal{R}^2) \times \mathcal{B}(\mathcal{R}) \to [0, 1]$ be such that $S_n(B, C) = P(\mathbf{x}_1 \in B,\ h_n \in C)$ $\forall B \in \mathcal{B}(\mathcal{R}^2)$, $C \in \mathcal{B}(\mathcal{R})$ and $\forall n \geq 2$. Let $\mu_{1n}(B, \alpha) = P(\mathbf{x}_1 \in B / h_n = \alpha)$ $\forall \alpha$ and $\forall n \geq 2$. Let ε be a small positive real number. Let the sequence of sets A_n be such that $P(h_n \in A_n) \geq 1 - \varepsilon$ $\forall n$. Let $\nu_{\xi n}(A_n) = P(h_n \in A_n / \mathbf{x}_1 = \xi)$. Let $b_n(\xi) = \nu_{\xi n}(A_n)$.

Assumption 2. Let there exists $M_3 > 0$ such that $\nu_{\xi n}(A_n) \geq 1 - \varepsilon$ $\forall n > M_3$ and $\forall \xi$ and for every such sequence A_n.

Lemma 5. Let $g(\mathbf{x}_1, h_n) = (1/h_n^2)K(\mathbf{x} - \mathbf{x}_1/h_n)$.

Then

$$\int_{\Gamma_{n1}} g(\mathbf{x}_1, h_n)\, dS_n < \frac{\varepsilon M_1}{4} \quad \forall n \geq M_2$$

and

$$\int_{\Gamma_{n2}} g(\mathbf{x}_1, h_n)\, dS_n < \frac{\varepsilon M_1}{4} \quad \forall n \geq M_2$$

where

$$\Gamma_{n1} = (h_n \in [0, \gamma_n]) \cap (\mathbf{x}_1 \in \mathcal{R}^2)$$

and

$$\Gamma_{n2} = (h_n \in [t_n, \infty]) \cap (\mathbf{x}_1 \in \mathcal{R}^2).$$

Lemma 6. Let $b_{1n}(\xi, \mathbf{x}) = (1/t_n^2)K(\mathbf{x} - \xi/\gamma_n)f(\xi)$ and $b_{2n}(\xi, \mathbf{x}) = (1/\gamma_n^2)K(\mathbf{x} - \xi/t_n)f(\xi)$. Then for $\delta_2 > 0$ $\exists M_4 > 0$ such that:

$$\left|\int_{\mathcal{R}^2} b_{1n}(\xi, \mathbf{x}) b_n(\xi)\, d\xi - \int_{\mathcal{R}^2} b_{1n}(\xi, \mathbf{x})\, d\xi\right| \leq \varepsilon(f(\mathbf{x}) - \delta_2)$$
$$\forall n \geq Max(M_3, M_4)$$

and also for $\delta > 0$ $\exists M_5 > 0$ such that:

$$\left|\int_{\mathcal{R}^2} b_{2n}(\xi, \mathbf{x}) b_n(\xi)\, d\xi - \int_{\mathcal{R}^2} b_{2n}(\xi, \mathbf{x})\, d\xi\right| \leq \varepsilon(f(\mathbf{x}) + \delta_2)$$
$$\forall n \geq Max(M_3, M_5)$$

Proofs of Lemmas 5 and 6 are given in the Appendix.

Theorem 1 (asymptotic unbiasedness). $E[f_n(\mathbf{x})] \to f(\mathbf{x})$ for every continuity point $\mathbf{x}$ of f in $\mathcal{R}^2$.

Proof.

$$E[f_n(\mathbf{x})] = \frac{1}{n} E\left[\frac{1}{h_n^2}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right]$$
$$= E\left[\frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_1}{h_n}\right)\right],$$
$$\text{(by Lemma 3)} = \int_0^{\infty}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n$$
$$= \int_0^{\gamma_n}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n + \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n$$
$$+ \int_{t_n}^{\infty}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n$$
$$= I_{1n} + I_{2n} + I_{3n}$$

where

$$I_{1n} = \int_0^{\gamma_n}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n$$

$$I_{2n} = \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n$$

and

$$I_{3n} = \int_{t_n}^{\infty}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n.$$

Now both I_{1n} and I_{3n} are less than $(M_1\varepsilon/4)$ $\forall n \geq M_2$, by Lemma 5. In $A_n = (t_n \geq h_n \geq \gamma_n)$:

$$\frac{1}{t_n^2} \leq \frac{1}{h_n^2} \leq \frac{1}{\gamma_n^2}$$
$$\Rightarrow \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{t_n^2}K\left(\frac{\mathbf{x}-\xi}{\gamma_n}\right) dS_n$$
$$\leq \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\xi}{h_n}\right) dS_n$$
$$\leq \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{\gamma_n^2}K\left(\frac{\mathbf{x}-\xi}{t_n}\right) dS_n$$
$$\Rightarrow I_{4n} \leq I_{2n} \leq I_{5n}$$

where

$$I_{4n} = \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{t_n^2}K\left(\frac{\mathbf{x}-\xi}{\gamma_n}\right) dS_n$$

and

$$I_{5n} = \int_{\gamma_n}^{t_n}\int_{\mathcal{R}^2} \frac{1}{\gamma_n^2}K\left(\frac{\mathbf{x}-\xi}{\gamma_n}\right) dS_n.$$

So $I_{4n} - f(\mathbf{x}) \leq I_{2n} - f(\mathbf{x}) \leq I_{5n} - f(\mathbf{x})$.

Now

$$|I_{2n}-f(\mathbf{x})| \geq Min\{|I_{4n}-f(\mathbf{x})|, |I_{5n}-f(\mathbf{x})|\}$$

and

$$|I_{2n}-f(\mathbf{x})| \leq Max\{|I_{4n}-f(\mathbf{x})|, |I_{5n}-f(\mathbf{x})|\}.$$

Now

$|I_{4n}-f(\mathbf{x})| \leq \varepsilon(f(\mathbf{x})+\delta_2) \quad \forall n \geq Max[M_3, M_4]$, from Lemma 6

and $|I_{5n}-f(\mathbf{x})| \leq \varepsilon(f(\mathbf{x})-\delta_2) \quad \forall n \geq Max[M_3, M_5]$, from Lemma 6.

Thus

$$|I_{2n}-f(\mathbf{x})| \leq \varepsilon(f(\mathbf{x})-\delta_2) \quad \forall n \geq Max[M_3, M_4, M_5].$$

Now

$$E[f_n(\mathbf{x})]-f(\mathbf{x}) = I_{1n}+I_{2n}-f(\mathbf{x})+I_{3n}.$$

Therefore,

$$|E[f_n(\mathbf{x})]-f(\mathbf{x})| \leq |I_{1n}|-|I_{2n}-f(\mathbf{x})|+|I_{3n}|.$$

Let us choose $N_1 = Max\{M_2, M_3, M_4, M_5\}$.

Thus,

$$|E[f_n(\mathbf{x})]-f(\mathbf{x})| \leq (\varepsilon M_1/4)+\varepsilon(f(\mathbf{x})+\delta_2)+(\varepsilon M_1/4) \quad \forall n \geq N_1$$

$$= \varepsilon((M_1/2)+f(\mathbf{x})+\delta_2) \quad \forall n \geq N_1$$

Since $(M_1/2)+f(\mathbf{x})+\delta_2$ is a finite quantity and ε is very small positive quantity, we have:

$$E[f_n(\mathbf{x})] \to f(\mathbf{x}) \text{ as } n \to \infty.$$

Hence the theorem. □

Let us define:

$$Y_{in}(\mathbf{x}) = \frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) - E\left[\frac{1}{nh_n^2}\sum_{j=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\right]$$

and

$$U_n = \frac{1}{n^2}E[\textstyle\sum_{i\neq j} Y_{in}(\mathbf{x})Y_{jn}(\mathbf{x})].$$

Let us define:

$$\sup_{\alpha\in(0,\infty)} \frac{g_n(\alpha)}{n\alpha^2} = d_n,$$

where $g_n(\alpha)$ is the density of h_n.

Assumption 3. $U_n \to 0$ as $n \to \infty$.

Assumption 4. $\text{Sup}\{d_n: n = 1, 2, \ldots\} < M_6 < \infty$.

Theorem 2. (Theorem of consistency)

$$E[f(\mathbf{x})-E[f_n(\mathbf{x})]]^2 \to 0 \text{ as } n \to \infty.$$

Proof.

$$E[f_n(\mathbf{x})-E[f_n(\mathbf{x})]]^2$$

$$= E\left[\frac{1}{nh_n^2}\sum_{i=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) - E\left[\frac{1}{nh_n^2}\sum_{i=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\right]\right]^2$$

$$= \frac{1}{n^2}E\left[\sum_{i=1}^{n}Y_{in}^2(\mathbf{x}) + \sum_{i\neq j}Y_{in}(\mathbf{x})Y_{jn}(\mathbf{x})\right]$$

$$= \frac{1}{n}E[Y_{1n}^2(\mathbf{x})] + U_n. \tag{9}$$

From Assumption 3, equation (9) becomes:

$$E[f_n(\mathbf{x})-E[f_n(\mathbf{x})]]^2 = \frac{1}{n}E\left[\frac{1}{h_n^4}K^2\left(\frac{\mathbf{x}-\mathbf{x}_1}{h_n}\right)\right] - \frac{1}{n}E^2[f_n(\mathbf{x})]. \tag{10}$$

As $n \to \infty$ the second term of equation (10) will be zero, because $E^2[f_n(\mathbf{x})]$ is finite. Now taking limit as $n \to \infty$ on both sides of equation (10) then we have:

$$\lim_{n\to\infty} E[f_n(\mathbf{x})-E[f_n(\mathbf{x})]]^2$$

$$= \lim_{n\to\infty}\frac{1}{n}E\left[\frac{1}{h_n^4}K^2\left(\frac{\mathbf{x}-\mathbf{x}_1}{h_n}\right)\right]$$

$$= \lim_{n\to\infty}\int_0^\infty\int_{\mathscr{R}^q}\frac{1}{n\alpha^4}K^2\left(\frac{\mathbf{x}-\xi}{\alpha}\right)\mathrm{d}S_n$$

$$= \lim_{n\to\infty}\int_0^\infty\left[\int_{\mathscr{R}^q}\frac{1}{n\alpha^4}K^2\left(\frac{\mathbf{x}-\xi}{\alpha}\right)\mu_{1n}(\mathrm{d}\xi,\alpha)\right]\mathrm{d}\beta_n(\alpha)$$

$$= \lim_{n\to\infty}\int_0^\infty\frac{1}{16n\alpha^4}\mu_{1n}[\mathscr{N}(\alpha),\alpha]\,\mathrm{d}\beta_n(\alpha)$$

$$\leq \lim_{n\to\infty}\int_0^\infty\frac{1}{16n\alpha^4}\alpha^2M_1\,\mathrm{d}\beta_n(\alpha) \quad \forall n \geq M_2$$

$$\leq \lim_{n\to\infty}\frac{M_1}{16}E\left[\frac{1}{nh_n^2}\right] = \lim_{n\to\infty}\int_0^\infty\frac{1}{n\alpha^2}g_n(\alpha)\,\mathrm{d}\alpha$$

$$\leq \lim_{n\to\infty}\frac{M_1}{16}\left[\int_0^{\gamma_n}\frac{1}{n\alpha^2}g_n(\alpha)\,\mathrm{d}\alpha + \int_{\gamma_n}^\infty\frac{1}{n\alpha^2}g_n(\alpha)\,\mathrm{d}\alpha\right]. \tag{11}$$

Thus from Assumption 4, equation (11) becomes:

$$\lim_{n\to\infty} E[f_n(\mathbf{x})-E[f_n(\mathbf{x})]]^2$$

$$\leq \lim_{n\to\infty}\frac{M_1}{16}\left[E\left[\frac{1}{nh_n^2}\right] \leq M_6\int_0^{\gamma_n}\mathrm{d}\alpha + \frac{1}{n\gamma_n^2}P(h_n \geq \gamma_n)\right]$$

$$\leq \lim_{n\to\infty}\frac{M_1}{16}\left[M_6\gamma_n + \frac{1}{n\gamma_n^2}\cdot 1\right] = 0$$

(because $0 < M_6 < \infty$ and $\gamma_n \to 0$ and $n\gamma_n^2 \to \infty$ as $n \to \infty$). □

3.2. *Generalization to higher dimension* ($q \geq 3$)

Let f, A, $\mathscr{M}$, $\mathscr{C}_A$ be now defined on $\mathscr{R}^q$, $q \geq 3$ under the same conditions as in Section 3.1.

Let $\mathbf{x}_i$, $i = 1$, n be now q-dimensional vectors with density f.

Again, f_n and h_n are defined by equations (2) and (3). Let us define $\mathscr{N}_x(\alpha) = \{\mathbf{y} = (y_1, y_2, \ldots, y_q)': |x_i - y_i| \leq \alpha, \forall i = 1, 2, \ldots, q\}$, where $\mathbf{x} = (x_1, x_2, \ldots, x_q)'$. Also, let S_n: $\mathscr{B}(\mathscr{R}^q) \times \mathscr{B}(\mathscr{R}) \to [0, 1]$ be such that $S_n(B, C) = P(\mathbf{x}_1 \in B, h_n \in C) \ \forall B \in \mathscr{B}(\mathscr{R}^q)$ and $C \in \mathscr{B}(\mathscr{R})$. Let $\mu_{1n}(B, \alpha) = P(\mathbf{x}_1 \in B | h_n = \alpha) \ \forall \alpha$ and $\forall n \geq 2$. Let ε be a small positive real number. Let the sequence of sets A_n be such that $P(h_n \in A_n) \geq 1 - \varepsilon \ \forall n$. Let $\nu_{2n}(A_n) = P(h_n \in A_n | \mathbf{x}_1 = \xi)$. Let $b_n(\xi) = \nu_{2n}(A_n)$. In addition, let:

$$Y_{in}(\mathbf{x}) = \frac{1}{h_n^q}K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) - E\left[\frac{1}{nh_n^q}\sum_{j=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\right]$$

and $U_n = (1/n^2)E[\sum_{i\neq j} Y_{in}(\mathbf{x})Y_{jn}(\mathbf{x})]$.

Also, let:

$$\sup_{x\in(0,\infty)} \frac{g_n(x)}{nx^q} = d_{nq}$$

where $g_n(x)$ is the density of h_n.

In this case also, the assumptions required are identical to Assumptions 1–4 under the above symbolic definitions. Then, we have the following theorem

Theorem 3. $E[f_n(\mathbf{x})]\to f(\mathbf{x})$ and $E[f_n(\mathbf{x})-E[f_n(\mathbf{x})]]^2\to 0$ as $n\to\infty$ for every continuity point $\mathbf{x}$ of f in $\mathscr{R}^q$.

The proof is identical to that of Theorems 1 and 2 and has been omitted for brevity.

Note. In this paper, we have always considered $q\geq 2$. Note that, when $q=1$, l_n converges to a finite quantity since A is bounded. So l_n cannot go to ∞. Thus, the conditions of Parzen's theorem are not valid here. Due to this reason, we did not consider the case $q=1$.

4. EXPERIMENTAL RESULTS

The density estimation scheme described above has been implemented on various randomly generated data sets in space $\mathscr{R}^2$. In each case, at first, the set A representing the region where the density is defined has been chosen. Using a finite number n of samples and equations (2)–(4), the density function $f_n(\mathbf{x})$ is estimated. We have found MST with a simple algorithm whose computational complexity is $O(n^2)$, although the $O(n\log n)$ algorithm[8] is also available. The rest of the calculation for finding densities needs negligible computation.

Next, a set B of test points has been chosen where the estimated densities are evaluated. The test points belong to the interior, border and outside, but not very far from the border of A. Points outside the border are chosen to see how the estimated density behaves in the vicinity outside A. The points with a distance greater than h_n away from the border of A will have zero estimated density since contribution of $K(\mathbf{x})$ is zero there.

Two error functions are defined. One is the average of the sum of squares of the differences between the actual and estimated densities for the test set B, given by:

$$E_1 = \frac{1}{N}\sum_{\mathbf{x}\in B}[f_n(\mathbf{x})-f(\mathbf{x})]^2,$$

where N is the number of points in B. The other error function is the maximum absolute difference between the actual and estimated densities for the test points of B, given by:

$$E_2 = \operatorname*{Max}_{\mathbf{x}\in B}|f_n(\mathbf{x})-f(\mathbf{x})|.$$

For any good density estimation procedure we expect that E_1 and E_2 will decrease with increase in sample size, irrespective of underlying density.

We have successfully made many experiments with data sets having different densities.

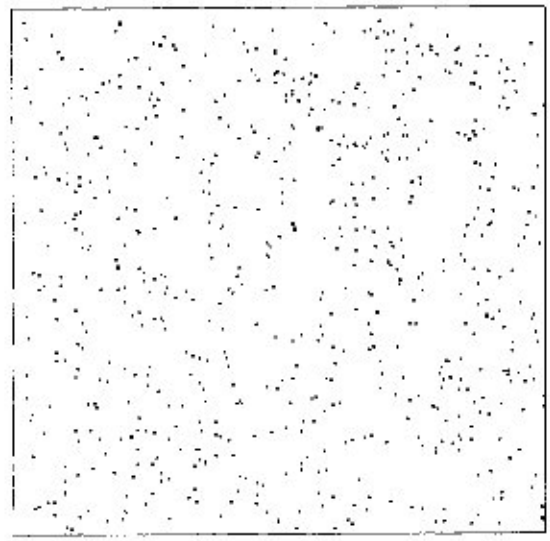


Fig. 1. The random sample of size 700 for Experiment 1.

Table 1. The decrease of error E_2 and the value of h_n for all data size of Experiments 1 and 2

	Experiment 1		Experiment 2	
n	h_n	E_2	h_n	E_2
100	0.7077	0.860235	1.2885	0.098872
200	0.6810	0.824802	1.2851	0.093398
300	0.7215	0.823908	1.3001	0.093230
400	0.6939	0.805295	1.2918	0.088180
500	0.7224	0.805192	1.2901	0.087782
600	0.7558	0.804987	1.2369	0.087623
700	0.7323	0.804782	1.2530	0.083667
800	0.7002	0.804756	1.2461	0.083123
900	0.6938	0.804632	1.2372	0.083016
1000	0.6534	0.804534	1.2511	0.082989
1500	0.5438	0.804251	1.2431	0.080687

Experiment 1 (uniform distribution over a square). For various values of the sample size n, the data has been generated randomly from $[0,1]\times[0,1]$ with uniform distribution. Thus, $A=[0,1]\times[0,1]$ and

$$f(\mathbf{x}) = \begin{cases} 1 & \forall \mathbf{x}\in A \\ 0 & \text{otherwise.} \end{cases}$$

Figure 1 shows a random sample of size 700 from the above density. Table 1 shows h_n and E_2 for various values of n showing that the error E_2 decrease with increase in n. Figure 2 shows that the decrease in E_1 with n is faster than the linear rate.

An image-like description of the results has been provided in Fig. 3(a)–(d). These figures give the gray-value representation of the density, where lesser density appears whiter. Figure 3(a) shows the original density (which is 1 in this case) in darkest (*black*) colour. Zero density, on the other hand, appears whitest. The difference between the actual density and

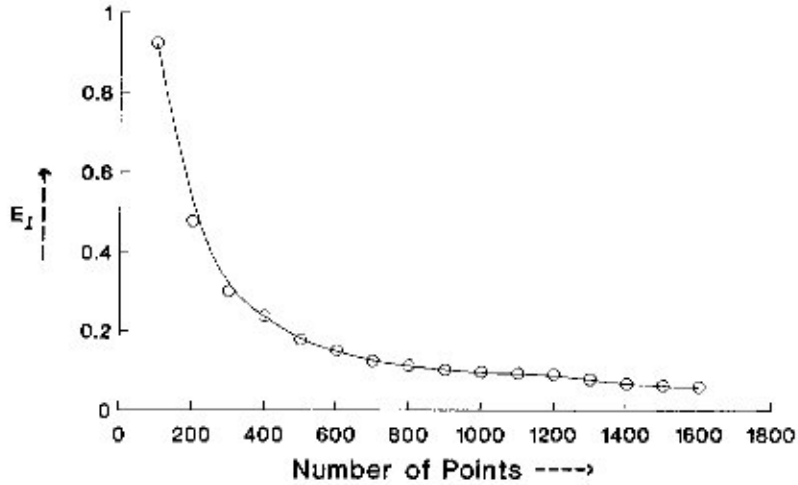


Fig. 2. The decrease of the error E_1 for Experiment 1.

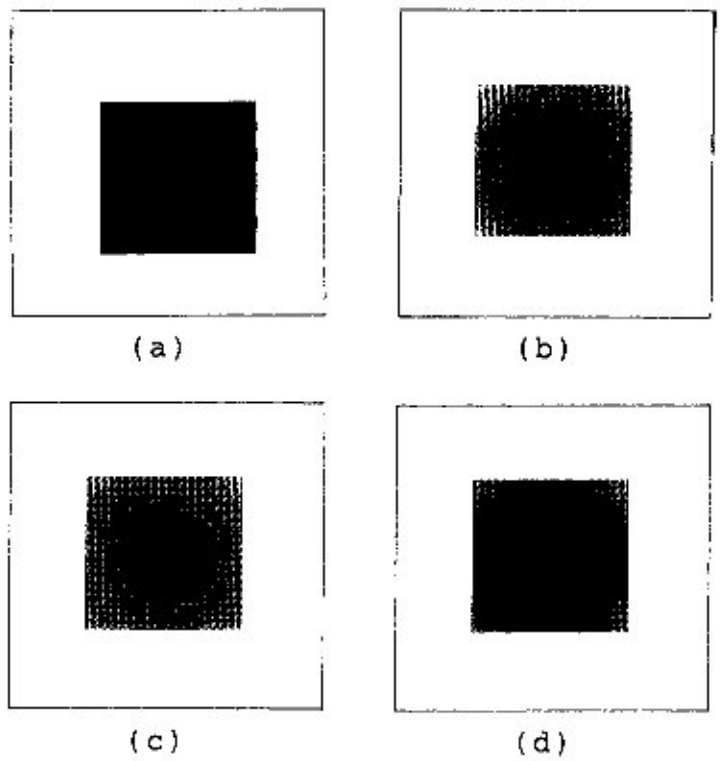


Fig. 3. The gray value representation of the density. (a) The original density with gray value 16. (b) The estimated density for 300 data. (c) The estimated density for 700 data. (d) The estimated density for 1500 data.

the estimated density has been scaled to 16 gray values, which are shown in Fig. 3(b)–(d) for $n = 300$, 700 and 1500, respectively. Note that the area of the *black portion* is increasing as the sample size n increases. Observe also that in Fig. 3(d) (i.e. $n = 1500$) the *black portion* occupies almost the entire area of the square.

It is understood that the kernel approach estimates non-zero density outside the defined space A, but it is desirable that the estimated value decreases sharply away from the border of A. To test the rate of decrease we considered the uniform density estimated by 1000 data and computed the average density outside the border of A. The size of the set is $[0, 1] \times [0, 1]$. Table 2 shows the estimated value against normal distance away for the border. Also, the estimated value is zero as the distance exceeds the window width. This table reflects the expected pattern of the estimated density away from the border.

Table 2. The estimated density away from the border

Distance	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Estimated average density	0.183	0.124	0.078	0.056	0.035	0.016	0.011

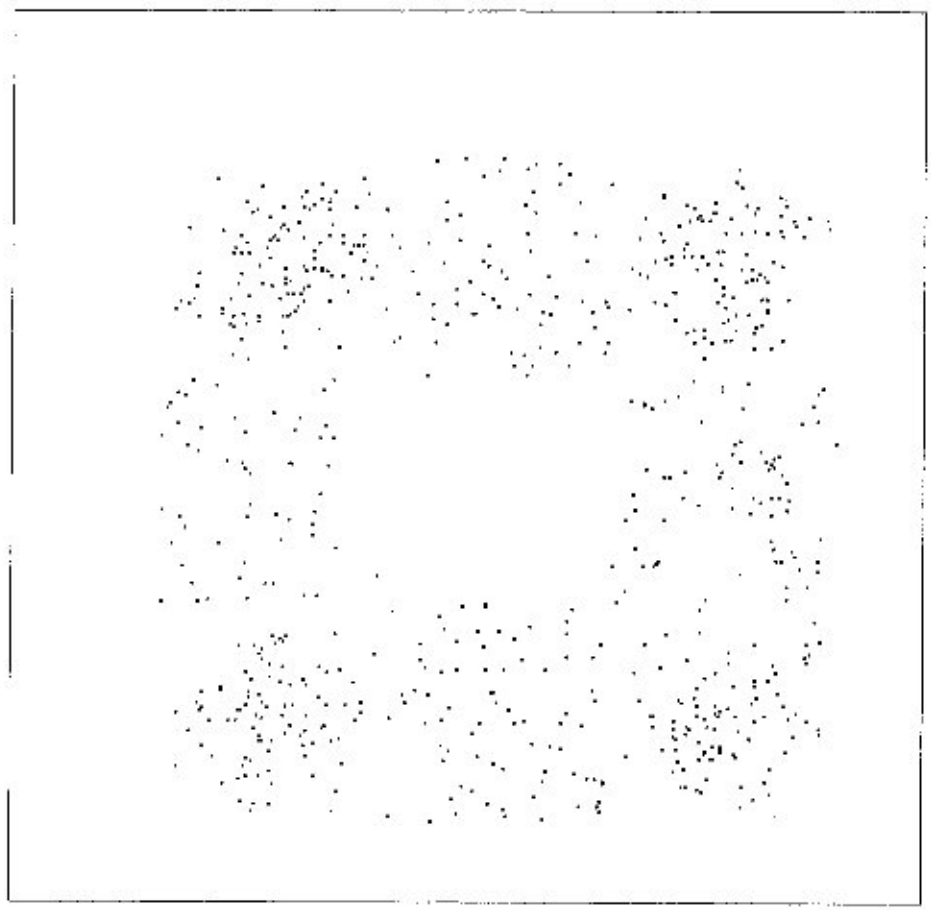


Fig. 4. The random sample of size 700 for Experiment 2.

Experiment 2 (multimodal density on a space with a hole). Here $A = A_1 \cup A_2 \cup \cdots \cup A_8$, where

$$A_1 = [0,1] \times [0,1], \quad A_2 = [0,1] \times [1,2],$$
$$A_3 = [0,1] \times [2,3], \quad A_4 = [1,2] \times [2,3],$$
$$A_5 = [2,3] \times [2,3], \quad A_6 = [2,3] \times [1,2],$$
$$A_7 = [2,3] \times [0,1] \quad \text{and} \quad A_8 = [1,2] \times [0,1].$$

We chose:

$$f(\mathbf{x}) = \begin{cases} 0.1 & \text{if } \mathbf{x} \in A_2 \cup A_4 \cup A_6 \cup A_8 \\ 0.1 + \Psi_1(x_1) \cdot \Psi_2(x_2) & \text{if } \mathbf{x} \in A_1 \cup A_3 \cup A_5 \cup A_7, \\ 0 & \text{otherwise} \end{cases}$$

where $\mathbf{x} = (x_1, x_2)'$,

$$v_i(x) = \begin{cases} \frac{1}{5}x & \text{if } 0 \le x \le 0.5 \\ \frac{1}{5}(1-x) & \text{if } 0.5 \le x \le 1 \end{cases}$$

and $\Psi_1(x_1) = v_1(x_1 - a)$ if $x_1 \in [a, a-1]$ $\forall a \in \mathscr{R}$ and $\Psi_2(x_2) = v_2(x_2 - a)$ if $x_2 \in [a, a+1]$ $\forall a \in \mathscr{R}$. Note that f is a mixture of uniform and triangular distributions.

Figure 4 shows a random sample of size 700 from the above density. Table 1 shows h_n and E_2 for each n, while Fig. 5 shows the decrease of E_1 with n.

Experiment 3 (data from Gaussian distribution). The experimental results stated earlier deal with distributions on bounded sets. To test how the approach work on unbounded set we considered data pooled from Gaussian distribution.

In the experiment the sample size n takes the values 100, 200, ..., 1000 and 1500. Three sets of data with zero mean and dispersion matrices $\Sigma_i = \begin{pmatrix} (\frac{1}{2})^{i-1} & 0 \\ 0 & (\frac{1}{2})^{i-1} \end{pmatrix}$ for $i = 1, 2, 3$ are generated. Table 3 shows h_n and E_2 for different cases.

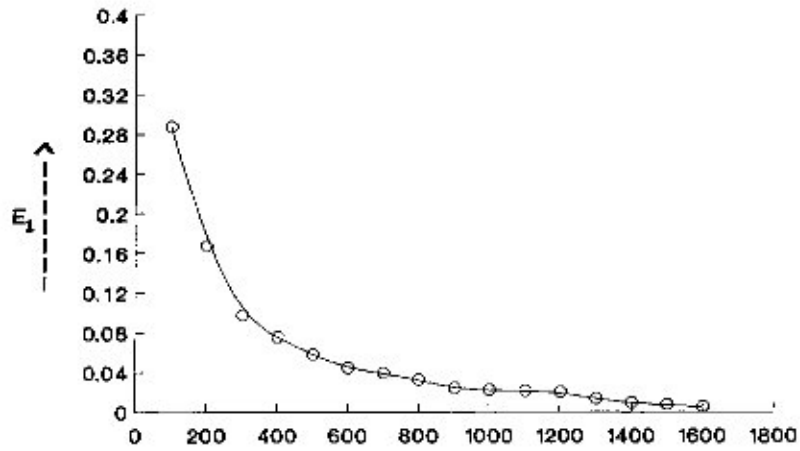


Fig. 5. The decrease of the error E_1 for Experiment 2.

Table 3. The decrease of error E_2 and the value of h_n for all data size from Gaussian distribution with mean (0,0) and dispersion matrices $\sum_1$, $\sum_2$ and $\sum_3$, respectively

	Experiment 3 ($\sum_1$)		Experiment 3 ($\sum_2$)		Experiment 3 ($\sum_3$)	
n	h_n	E_2	h_n	E_2	h_n	E_2
100	1.9737	0.161128	1.6597	0.143397	1.3956	0.116695
200	1.7964	0.158522	1.5106	0.138732	1.2702	0.105429
300	1.8999	0.158363	1.5976	0.138458	1.3434	0.105345
400	1.8144	0.157828	1.5257	0.135849	1.2830	0.104876
500	1.8344	1.157568	1.5425	0.135630	1.2971	0.104660
600	1.8668	0.157335	1.5658	0.135398	1.3200	0.104652
700	1.8460	0.156925	1.5587	0.135367	1.3314	0.104608
800	1.8132	0.156332	1.5379	0.134864	1.3048	0.104051
900	1.8316	0.156308	1.5421	0.134541	1.3150	0.103653
1000	1.7897	0.153075	1.5316	0.133277	1.2976	0.102353
1500	1.7323	0.152898	1.4989	0.133256	1.2638	0.100209

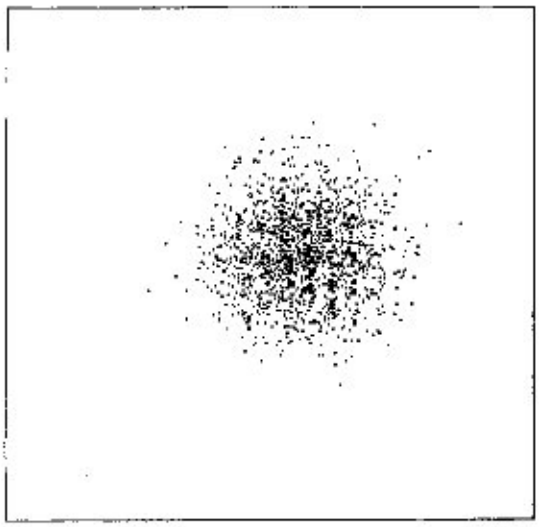


Fig. 6. The random sample of size 700 for Experiment 3.

Figure 6 show a random sample of size 700 from the above density with zero mean and dispersion matrices $\sum_1$. Figure 7 gives the graphical description of the decrease in the error value E_1 for different Gaussian distributions and for different sample sizes. Different markers [□, *, ○] are used to denote the error for different Gaussian distributions. Note that as i increases, the variances decrease and intuitively the density estimates should be more accurate. Figure 7 provides a demonstration of this convergence.

5. SOME APPLICATIONS

The proposed density estimation method is useful and important in a large class of problems of statistical pattern classification, clustering and interpretation of dot patterns. Here we present some newer applications involving dot patterns and data sets.

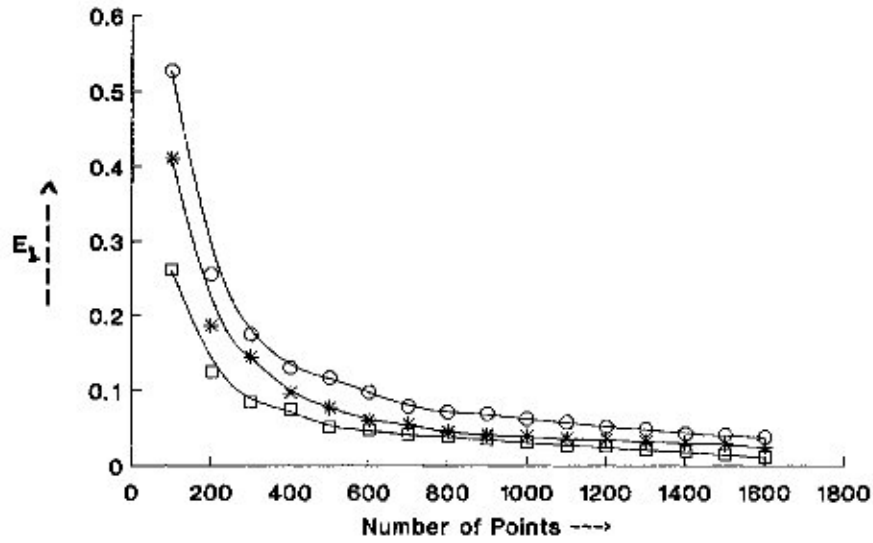


Fig. 7. The decrease of the error E_1 for different Gaussian distributions and for different sample sizes.

5.1. *Estimation of representative points*

Consider a set of objects represented as point data in $\mathcal{R}^n$ feature space. Given a set S of n data, we address the problem of selecting a small subset $V \subset S$ of $n_0 \ll n$ data that *faithfully* represents the spatial organization of original data. The solution to this problem can find applications in data compression, data clustering, pattern classification as well as statistical parameter estimation. For example, in clustering, many algorithms start with a few *seed* points, where each *seed* point represents the core of one cluster. The minimum distance classifier or the k-nearest neighbour classifier considers the *seed* points as the best patterns represen-

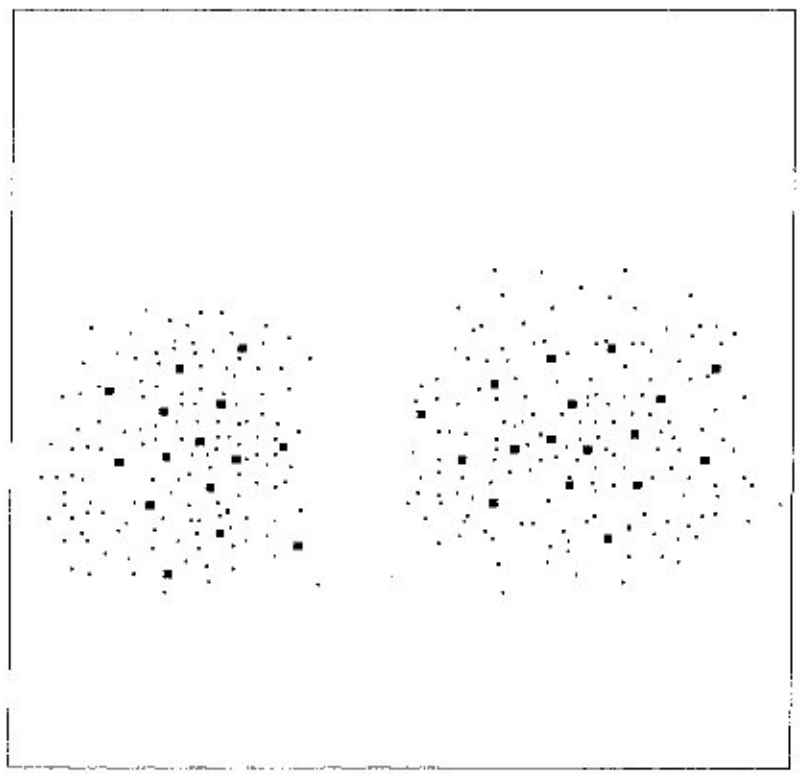


Fig. 8. Synthetic data showing two clusters and almost 10% representative points.

ting the class. Quite often, a single *best representative point* is assumed to be the *mode* of the pattern set. *Density* and *mode* estimation are two classical problems in statistics. In many situations, the problem of finding best representative points may be considered as a generalization of mode estimation and seed point detection problem.

The algorithm for obtaining *local best representation points* from $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathcal{R}^q$ is described below.

Algorithm RPD. Let the problem be that of choosing one representative out of k data units. Here we propose a *density* based method with the following steps.

Step 1. Compute the *density* for each datum $\mathbf{x}$ from the number of other data units within an open disc of radius h_n with $\mathbf{x}$ as the centre. At the point $\mathbf{x}_i$; $i = 1, 2, \ldots, n$.

Let $V_i = \{\mathbf{y}: \ \|\mathbf{x}_i - \mathbf{y}\| \le h_n, \mathbf{y} \in S\}, \quad i = 1, 2, \ldots, n.$

The density is defined as:

$$m_i = \frac{1}{2^q n h_n^q} \times \#V_i, \quad i = 1, 2, \ldots, n$$

($\#A$ is the number of points in the set A and q is the dimension of the space).

Step 2. Rearrange $m_1, m_2, \ldots, m_n$ in decreasing order. Let L be the ordered list. Let $i \leftarrow 1$.

Step 3. Choose the datum that tops the list L as the ith representative datum. If $i = n_0$ go to Step 6.

Step 4. Count the number of data in the current S. If the number is less than $k - 1$ then go to Step 6. Else, from the current S find the $k - 1$ nearest neighbours of the datum $\mathbf{x}$ which has been chosen in Step 3. Delete $\mathbf{x}$ and these $k - 1$ neighbours from L and S to obtain the list of L and S for the next iteration.

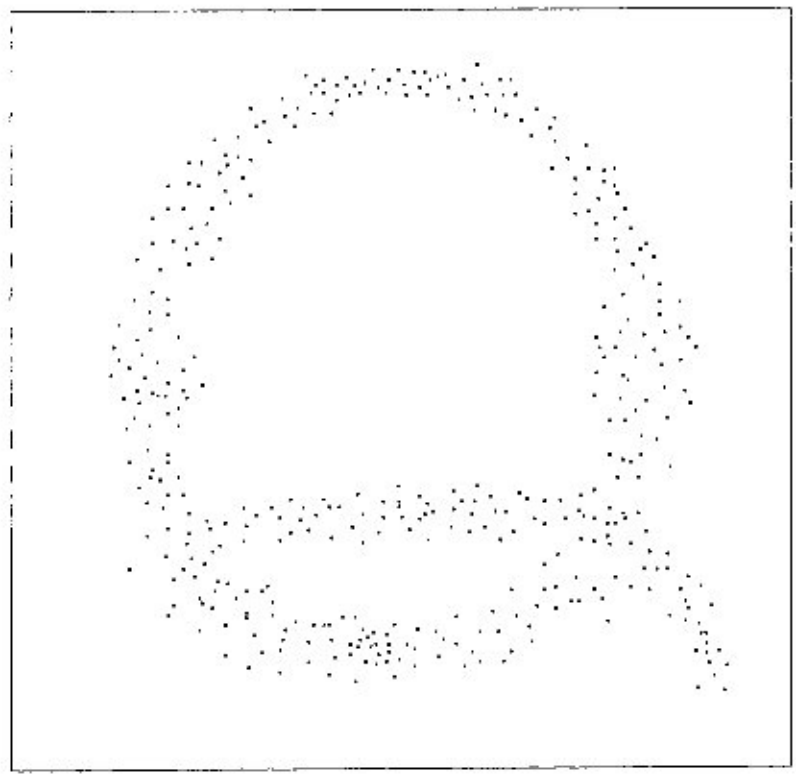
Step 5. Make $i \leftarrow i + 1$ and go to Step 3.

Step 6. Stop.

The representative point detection (RPD) algorithm, using the proposed density estimation method, can be applied for selecting a small subset (representative set) of the original data.

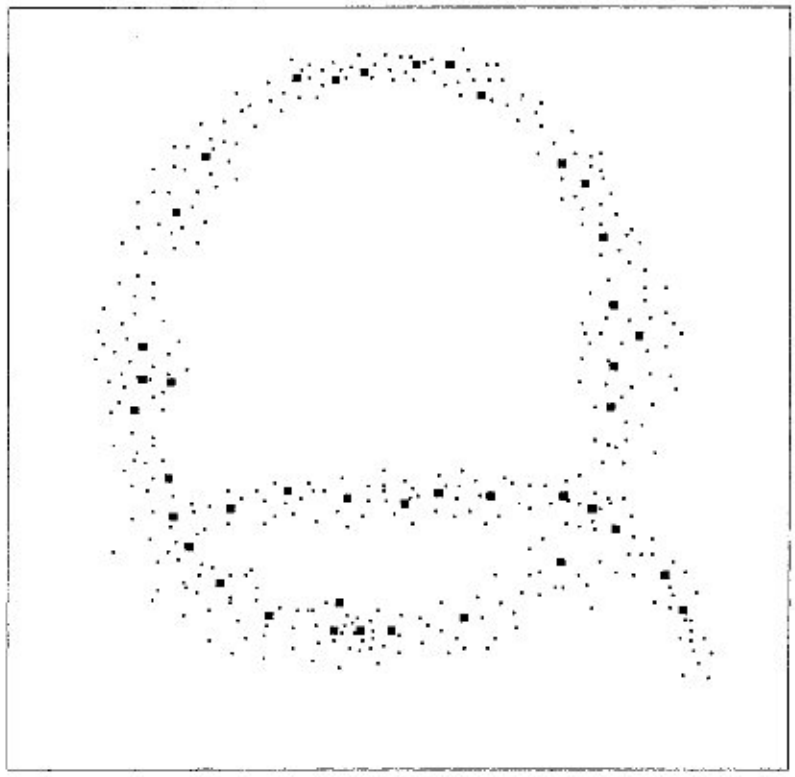
Figure 8 contains two prominent clusters where each cluster is compact and round in shape. We considered the problem of choosing nearly 10% representative points in a synthetic data of size 322 by using RPD algorithm. The result is shown in Fig. 8 and the representative points are marked by dark square.

The RPD algorithm is tested on point patterns of non-convex shapes as well. Figure 9(a) shows a Q-shaped data of size 412 with $h_n = 0.6534$. Almost 10% representative points are shown in Fig. 9(b).

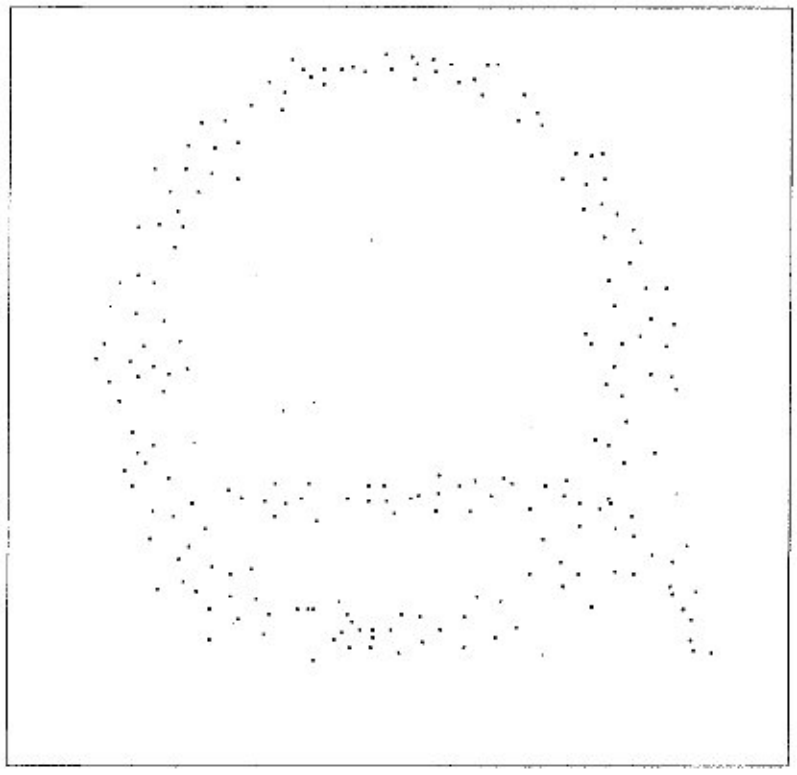


(a)

Fig. 9. An example of Q-shaped synthetic data. (a) A Q-shaped synthetic data of size 412. (b) Almost 10% representative points. (c) Almost 50% representative points. (d) Skeleton by rejecting 80% data. (e) Border points obtained by retaining 15% low density data.

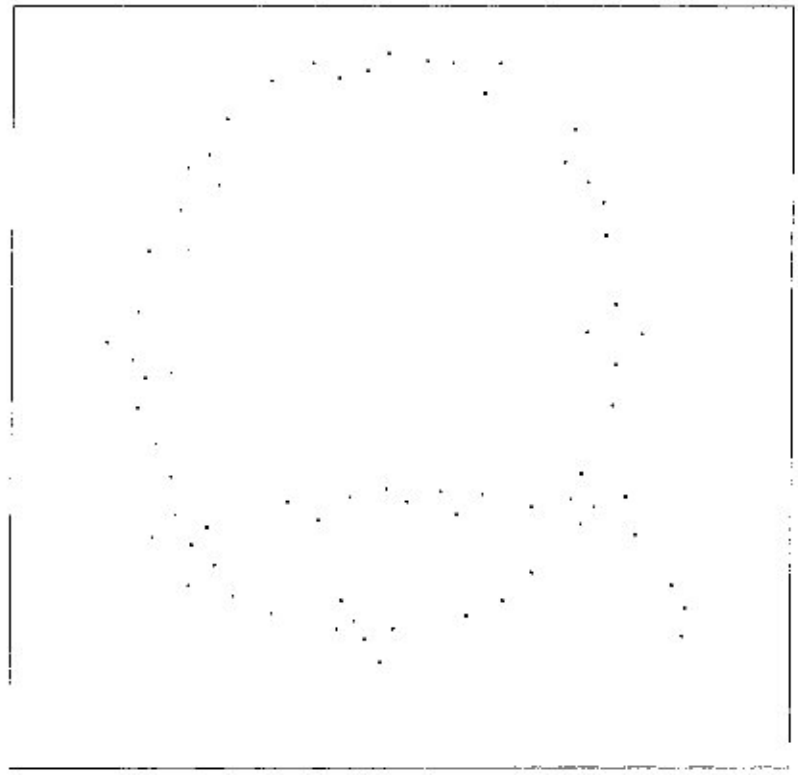


(b)

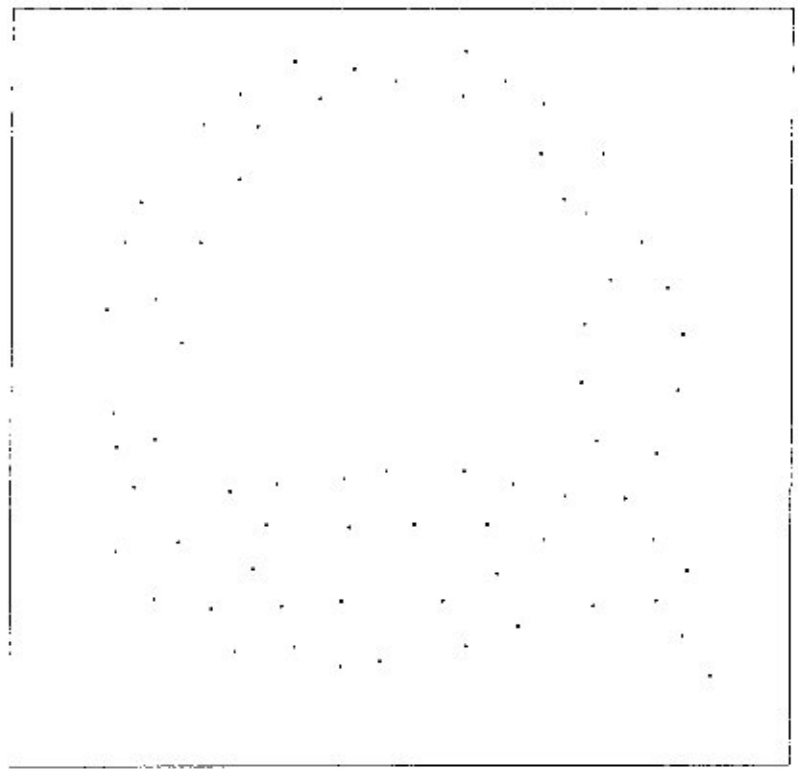


(c)

Fig. 9. (Continued).



(d)



(e)

Fig. 9. (Continued).

5.2. *Border point detection and dot pattern thinning*

In addition to representative points detection and seed point evaluation, the RPD can be applied for data reduction and outlier rejection. Detection of border points of a dot pattern as well as finding a *thinned* version of the dot pattern can also be carried out by such an algorithm. The basic idea is that the low density points are usually border points and the high density points are the skeleton of the dot pattern. So the border and thinned version of a dot pattern are mostly dependent on the value of the estimated density. The better results for border and thinned version of a dot pattern can be found by better approximation of the estimated density. In this paper we employed the RPD algorithm, using the proposed density estimation method, to find the border and thinned version of a dot pattern. Consider, for example, Fig. 9(a). By using the RPD algorithm, 50% of the data are rejected. It can be seen from Fig. 9(c) that the nature of data distribution appears identical and the shape of the data is retained. On the other hand, if 80% of the data are rejected, then we obtain a *thinned* version of the shape as in Fig. 9(d). Similarly, if we retain 15% of the data from the low density side through the RPD algorithm, we obtain the border points of the shape as Fig. 9(e) and all these results reflect the utility of the proposed density estimation method.

However, density alone cannot capture the notion of border points, because if the interior portion of any pattern is sparsely populated, then the interior points are also detected as border points. We have a perceptual notion about the points lying on the border compared with those of the interior of the data set. Border points are not surrounded by other points in all directions while the interior points are. The present approach of border point detection is based on this observation.[34,35]

Definition 1. A point $\mathbf{x} \in S$ is said to be an *opposite point* of $\mathbf{y} \in S$ with respect to $\mathbf{z} \in S$ if $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{y}$ almost lie in a straight line, i.e. if:

$$I(\mathbf{x}, \mathbf{y})_z = \frac{d(\mathbf{x}, \mathbf{y})}{d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})} \approx 1,$$

where $d(\mathbf{x}, \mathbf{y})$ is the Euclidean distance between two points $\mathbf{x}$ and $\mathbf{y}$.

Figure 10 shows that $(\mathbf{x}, \mathbf{y})$ are nearly opposite points with respect to $\mathbf{z}$.

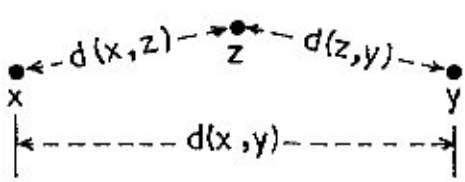


Fig. 10. Nearly opposite points with respect to a fixed point.

Note that if $\mathbf{x}$ is an opposite point of $\mathbf{y}$, then $\mathbf{y}$ is also an opposite point of $\mathbf{x}$ with respect to $\mathbf{z}$. $I(\mathbf{x}, \mathbf{y})_z$ may be termed the *degree of oppositeness* of $\mathbf{x}$ and $\mathbf{y}$ with respect to $\mathbf{z}$.

Consider a neighbourhood D around $\mathbf{z}$. Let I be the average of $I(\mathbf{x}, \mathbf{y})_z$; $\mathbf{x}, \mathbf{y} \in D$. If I is not approximately equal to 1, then $\mathbf{z}$ should be a corner point in the neighbourhood.

Definition 2. A point $\mathbf{x} \in S$ is said to be a *border point* if the value of the degree of oppositeness of $\mathbf{x}$ of k-neighbour $\ll k/2$.

Definition 3. A point $\mathbf{x} \in S$ is said to be an *interior point* if the value of the degree of oppositeness of $\mathbf{x}$ of k-neighbour $\approx k/2$.

Next, the border point detection algorithm of a data set $S = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subseteq \mathscr{R}^q$ is given below.

Algorithm BPD.

Step 1. Compute the *density* for each datum $\mathbf{x}$ from the number of other data units within an open disc of radius h_x with $\mathbf{x}$ as the centre. At the point $\mathbf{x}_i$; $i = 1, 2, \ldots, n$.

Let $V_i = \{\mathbf{y}: \|\mathbf{x}_i - \mathbf{y}\| \leq h_x, \mathbf{y} \in S\}$, $i = 1, 2, \ldots, n$.

The density is defined as:

$$m_i = \frac{1}{2^q n h_x^q} \times \#V_i, \quad i = 1, 2, \ldots, n.$$

Step 2. Rearrange $m_1, m_2, \ldots, m_n$ in decreasing order. Let L be the ordered list.

Step 3. Delete $[w_1\% \times n] = n_1$ data from the top of the list L, where '$[a]$' is the largest integer $\leq a$. Let L_1 be the set after deleting n_1 number of points from L and let $n_2 = n - n_1$.

Step 4. Compute $m_0 = [w_2\% \times n_2]$.

Step 5. Find the value of the degree of oppositeness of each point of L_1 from the original data set S with k-neighbour.

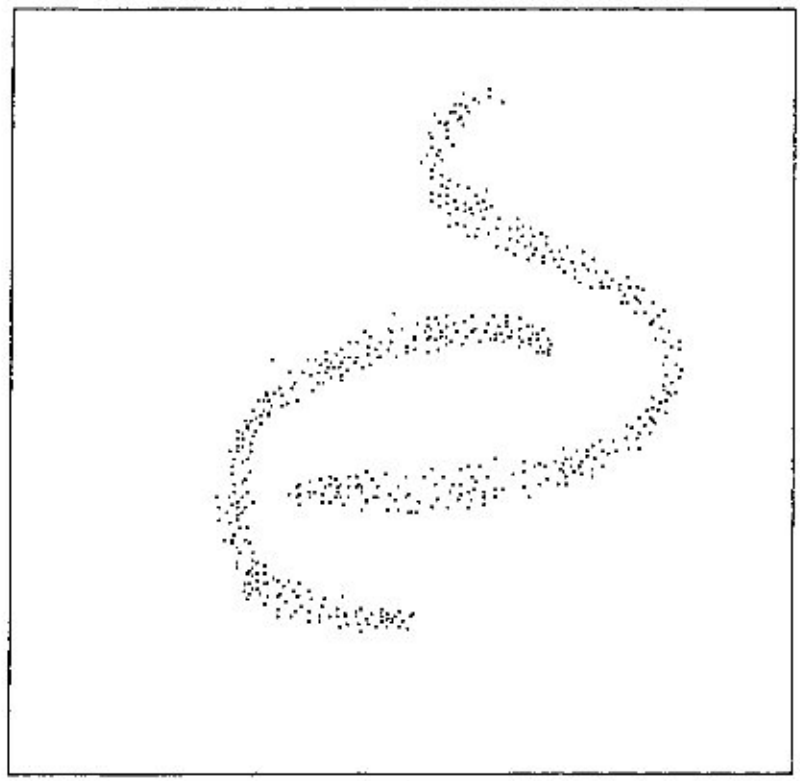
Step 6. Rearrange the points according to the increasing order of their value of the degree of oppositeness.

Step 7. Declare the first m_0 ranking points as m_0 border points.

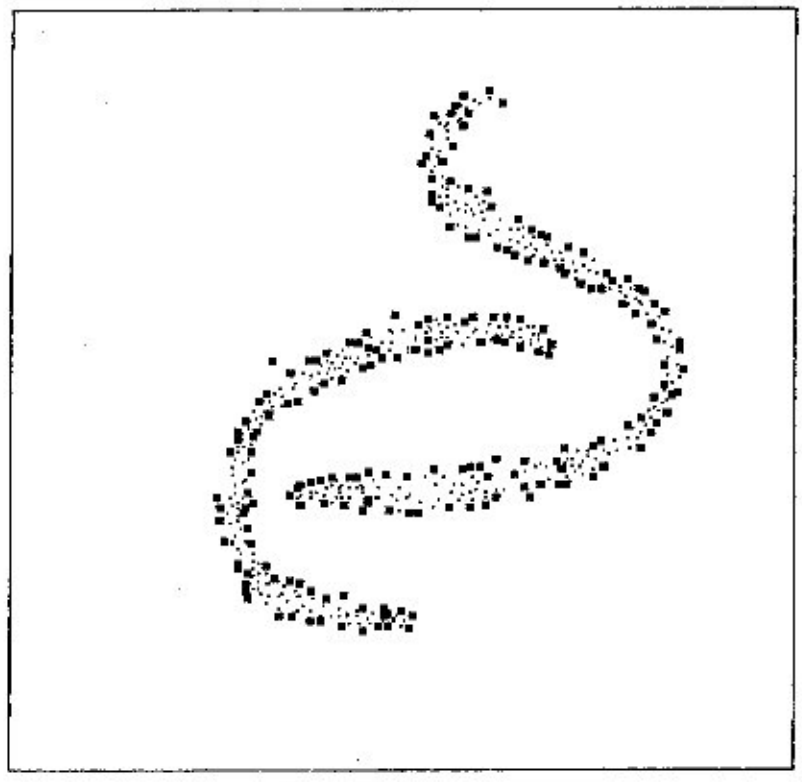
To test the efficiency of the border point detection (BPD) algorithm, several 2-D data were generated. Figure 11(a) shows a non-convex shape data. The border points of this data are marked by dark small square [Fig. 11(b)].

Step 3 is used for deleting those data whose densities are most significant, because the data with most significant density will be the interior point. In our case we chose $w_1 = 40$.

The BPD algorithm is also useful for the basic idea about the shape of the dot pattern. If the dot pattern is of hyperspherical shape then the maximum and minimum of pairwise distances between the border points are almost equal. If the dot pattern is elongated or non-convex, then the difference between the maximum and minimum of pairwise distances is usually greater than some threshold value.



(a)



(b)

Fig. 11. Non-convex shaped data. (a) A non-convex shaped data of size 570. (b) Border points obtained by applying the BPD algorithm.

6. DISCUSSIONS

A data-driven procedure of density estimation has been suggested in this paper. It is theoretically shown that under certain assumptions, densities on bounded sets can be consistently estimated using a kernel-based approach, where the width of window (i.e. h_n) is obtained from the minimal spanning tree of the observations. In addition to the results presented here we verified our estimation procedure on bounded sets with/without holes where the distributions are uniform, triangular (unimodal) or mixed (multimodal). Another experimental result supports the idea that the same procedure can be used for estimating Gaussian density also.

Some of the theoretical assumptions made in Section 3 can be probably relaxed giving rise to similar results. For example, even for unbounded sets, similar theoretical results may be shown. Note that many distributions (such as normal) follow the condition $\mathbf{x}f(\mathbf{x}) \to 0$ as $|\mathbf{x}| \to \infty$. It will be interesting to investigate whether the proposed estimator is consistent and asymptotically unbiased under this condition.

There exist numerous applications of density estimation technique. A wide variety of pattern recognition approaches employ density at some stage. We have considered some new problems where our proposed density estimation technique is useful.

7. SUMMARY

This paper concerns the estimation of probability density from a finite set of data points. One of the popular approaches of estimating density is due to Parzen and Cacoullos, where the density estimate at a point $\mathbf{x}$ is obtained by accumulating evidence from a window around $\mathbf{x}$. It can be proved that the density estimate converges to the true density for window of any size provided the data is unlimited. However, the user finds a problem with finite data set, because there exist no guidelines of choosing the window size in such a situation. Arbitrary choice of window size can lead to a practically useless estimate of density. For example, if the size is too small the estimate will suffer from too much of variability while a large window size will smoothen the density to an unacceptable limit.

In this paper we propose a kernel-based method where the window size is derived from the given set of data. It is understood that the window size should depend on (1) the number of data n and (2) their spatial arrangement. These two factors are combined by finding the minimal spanning tree of the data set and normalizing its length by the number of data n. More specifically the window width is proposed as $h_n = (l_n/n)^{1/q}$. Since h_n is a random variable as opposed to constant h_n in the original approach, the arguments for the proof of convergence due to Parzen and Cacoullos do not hold here. So, we have to prove the consistency and asymptotic unbiasedness of the estimator with our proposed h_n. The theoretical development initially carried out on 2-D has been extended to space of arbitrary dimensions. A few assumptions were used to arrive at the proof of convergence. The main assumption involves boundedness and path connectedness of the set.

The density estimation scheme described here has been experimentally tested on various data sets. The data are drawn from a source of known distribution or mixture of distributions. The results show the expected behaviour of the estimated density. An image-like description of the results is also presented for one set of data to obtain a visual effect of the estimated density.

Since these data are on bounded sets, an ideal estimation procedure should show zero density outside the set boundary. However, any practical estimator will have a *spill-over* effect. This effect is small for the window size chosen by our approach.

Although the approach is theoretically established for bounded sets, we took data from an unbounded distribution, such as a Gaussian distribution, and experimentally found that our estimation procedure works well for them also. These results led us to believe that the approach could also be useful for density estimation over unbounded sets.

The representative point detection algorithm, using the proposed density estimation procedure, is also applied for data reduction and outlier rejection. Detection of border points of a dot pattern and a thinned version of the dot pattern are also be found by such an algorithm.

Acknowledgement—The authors express their gratitude to Professor A. K. Jain, Michigan State University for his encouragement and constructive suggestions during the preparation of the manuscript. Secretarial help rendered by Mr S. Chakraborty is acknowledged with thanks.

REFERENCES

1. V. Ryzin, On a histogram method of density estimation, *Commun. Statist.* **2**, 493–506 (1973).
2. L. I. Boneva, D. Kendall and I. Stefanov, Spline transformations, *J. R. Statist. Soc. Ser. B* **33**, 1–70 (1971).
3. E. Parzen, On the estimation of a probability density function and the mode. *Ann. Math. Statist.* **33**, 1065–1076 (1962).
4. T. Cacoullos, Estimation of a multivariate density, *Ann. Inst. Statist. Math.* **18**, 178–189 (1966).
5. D. O. Loftsgaarden and C. P. Quensenberry, A nonparametric estimate of a multivariate density function, *Ann. Math. Statist.* **36**, 1049–1051 (1965).
6. K. Fukunaga and D. Hosteller, Optimization of *k*-nearest neighbor density estimates, *IEEE Trans. Inform. Theory* **IT-19**, 320–336 (1973).
7. N. N. Cencov, Evaluation of an unknown distribution density from observations, *Soviet Math.* **3**, 1559–1562 (1962).
8. J. B. Kruskal, On the shortest spanning subtree of a graph and the travelling salesman problem, *Proc. Am. Math. Soc.* **7**, 48–50 (1956).
9. B. B. Winter, Rate of strong consistency of two nonparametric density estimators. *Ann. Statist.* **3**, 759–766 (1975).

10. I. J. Good and R. A. Gaskins, Nonparametric roughness penalties for probability densities, *Biometrika* **58**, 255–277 (1971).
11. R. L. Kashyap and C. C. Blaydon, Estimation of probability density and distribution function, *IEEE Trans. Inform. Theory* **IT-14**, 459–466 (1968).
12. W. Wertz, *Statistical Density Estimation - a Survey*. Vandenhoeck and Ruprecht, Gottingen (1978).
13. B. L. S. Prakasa Rao, *Nonparametric Functional Estimation*. Academic Press, New York (1983).
14. C. A. Murthy, On consistent estimation of classes in $\mathscr{R}^2$ in the context of cluster analysis. Ph.D. Thesis, I.S.I., Calcutta (1989).
15. P. Billingsley, *Probability and Measure*. John Wiley and Sons, New York (1979).
16. C. T. Zahn, Graph theoretic methods for detecting and describing gestalt clusters, *IEEE Trans. Compu.* **20**, 68–86 (1971).
17. K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, New York (1972).
18. A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, New Jersey (1988).
19. D. Chaudhuri, C. A. Murthy and B. B. Chaudhuri, Finding a subset of representative points in a data set, *IEEE Trans. Syst. Man Cybernet.* **24**(9), 1416–1424 (1994).
20. R. P. W. Duin, On the choice of smoothing j parameters for Parzen estimators of probability density functions, *IEEE Trans. Comput.* **C-25**, 1175–1179 (1976).
21. W. L. G. Koontz and K. Fukunaga, Asymptotic analysis of a nonparametric clustering technique, *IEEE Trans. Comput.* **C-21**, 967–974 (September 1972).
22. S. J. Sheather, A data-based algorithm for choosing the window width when estimating the density at a point, *Comput. Statist. Data Anal.* **1**, 229–238 (1983).
23. S. J. Sheather, An improved data-based algorithm for choosing the window width when estimating the density at a point, *Comput. Statist. Data Anal.* **4**, 61–65 (1986).
24. W. Greblicki, Pattern recognition procedures with nonparametric density estimates, *IEEE Trans. Syst. Man Cybern.* **SMC-8**, 211–228 (1988).
25. A. K. Jain and M. D. Ramaswami, Classifier design with Parzen windows, *Pattern Recognition Artif. Intell.* 211–228 (1988).
26. R. Pyke, Spacings, *J. R. Statist. Soc. Ser. B.* **27**, 395–436 (1965).
27. L. Devroye, Laws of the iterated logarithm for order statistics of uniform spacings, *Ann. Probabil.* **9**(5), 860–867 (1981).
28. J. M. Steele, Growth rates of Euclidean minimal spanning trees with power weighted edges, *Ann. Probabil.* **16**(4), 1767–1787 (1988).
29. J. M. Steele, Probabilistic and worst case analysis of classical problems of combinatorial optimization in Euclidean space, *Math. Operat. Res.* **15**(4), 749–770 (1990).
30. J. M. Steele, Probabilistic algorithm for the directed traveling salesman problem, *Math. Operat. Res.* **11**(2), 343–350 (1986).
31. J. M. Steele, Complete convergence of short paths and Karp's algorithm for the TSP, *Math. Operat. Res.* **6**(3), 374–378 (1981).
32. J. Beardwood, J. H. Halton and J. M. Hammersley, The shortest path through many points, *Proc. Cambridge Philos. Soc.* **55**, 299–327 (1959).
33. J. H. Halton and R. Terada, A fast algorithm for the Euclidean traveling salesman problem, optimal with probability one, *SIAM. J. Comput.* **11**(1), 28–46 (1982).
34. B. B. Chaudhuri and D. Chaudhuri, Detection of representative points from a data set, Technical Report, TR/KBCS/2/93, Knowledge Based Computing Systems, Indian Statistical Institute, Calcutta.
35. D. Chaudhuri, Some studies on density estimation and data clustering techniques. Ph.D. Thesis, ISI, Calcutta (1994).

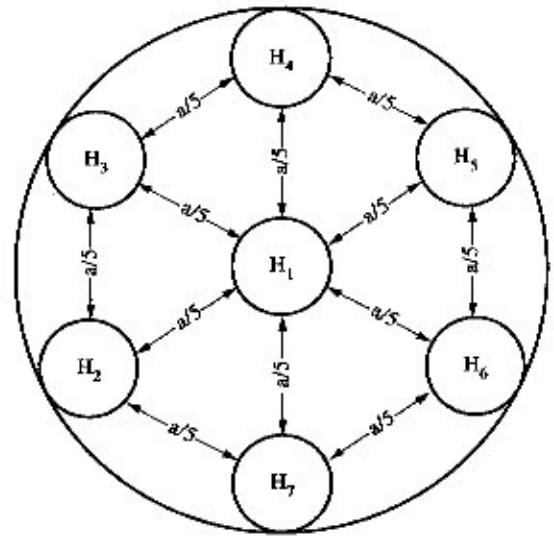


Fig. 12. Seven open discs of diameter $5a$.

APPENDIX

Proof of Lemma 1

Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be independent and identically distributed random vectors with density $f(\mathbf{x})$, where $f(\mathbf{x}) > 0\ \forall \mathbf{x} \in Int(A)$. Hence $Int(A) \neq \varnothing$.

Let $\eta \subseteq A$ and η is an open disc of diameter $5a$, $a > 0$. Draw $H_1, H_2, \ldots, H_7$ in η as shown in Fig. 12. $P(\mathbf{x}_1 \notin H_1, \mathbf{x}_2 \notin H_1, \ldots, \mathbf{x}_n \notin H_1) \to 0$ as $n \to \infty$ since $P(H_1) > 0$. $[P(H_1) = \int_{H_1} f(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} > 0$ since $f(\mathbf{x}) > 0\ \forall \mathbf{x} \in H_1]$.

So $P(\exists n$ such that $\mathbf{x}_n \in H_1) \to 1$ as $n \to \infty$.

Similarly $P(\exists n_1, n_2, \ldots, n_7$ such that $\mathbf{x}_{n_i} \in H_i) \to 1$ as $n \to \infty$. i.e. $P(l_n > 6a) \to 1$ as $n \to \infty$.

Similarly $P(l_n > 6a + 7.6 \cdot \frac{a}{5}) \to 1$ as $n \to \infty$. [by drawing 7 discs in each one of $H_1, H_2, \ldots, H_7$].

Again, applying the above process:

$$P\left(l_n > 6a + 7.6 \cdot \frac{a}{5} - 7^2 \cdot 6 \cdot \frac{a}{5^2}\right) \to 1 \text{ as } n \to \infty.$$

Note that

$$6a - 7.6 \cdot \frac{a}{5} + 7^2 \cdot 6 \cdot \frac{a}{5^2} - \cdots = 6a\left[1 + \frac{7}{5} + \left(\frac{7}{5}\right)^2 + \cdots\right] \to \infty.$$

Thus, $P(l_n > M) \to 1$ as $n \to \infty$ for every $M > 0$.

Therefore, $nh_n^2 \to \infty$ in probability as $n \to \infty$.

Proof of Lemma 2.

It is known that A is a class. Hence, $Int(A) \neq \varnothing$.

Let m_n be the maximum of the $(n-1)$ edge weights of MST.

Now $\frac{l_n}{n} = \frac{l_n}{n-1} \times \frac{n-1}{n} \leq m_n \times \frac{n-1}{n} < m_n$.

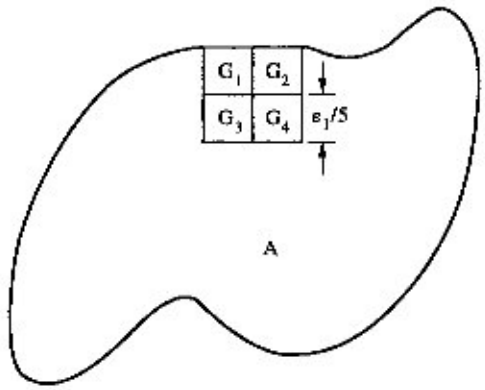
It suffices to show that $P(m_n < \varepsilon_1) \to 1$ as $n \to \infty$ for every $\varepsilon_1 > 0$.

Let $\varepsilon_1 > 0$. Cover the set A with open squares of size $(\varepsilon_1/5)$ so that union of these squares $\supseteq A$. Note that finitely many squares are sufficient to cover A since $\lambda(A) < \infty$ and A is bounded [λ is the Lebesgue measure] [as Fig. 13(a)].

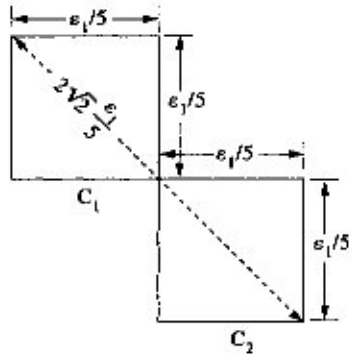
Let $G_1, G_2, \ldots, G_k$ be squares of size $(\varepsilon_1/5)$ such that:

$$\bigcup_{i=1}^{k} G_i \supseteq A \text{ and } G_i \cap A \neq \varnothing \quad \forall i = 1, 2, \ldots, k.$$

As $n \to \infty$, $P(\exists \mathbf{x}_i$ such that $\mathbf{x}_i \in G_1) \to 1$.



(a)



(b)

Fig. 13. (a) The set A with open squares of size $(\varepsilon_1/5)$. (b) Two consecutive squares C_1 and C_2.

Similarly $P(\exists n_1, n_2, \ldots, n_k$ such that $\mathbf{x}_{n_i} \in G_i$ $\forall i = 1, 2, \ldots, k) \to 1$ as $n \to \infty$. That means $P(n_n < \varepsilon_1) \to 1$ as $n \to \infty$ since for two consecutive squares C_1 and C_2 [as Fig. 13(b)]

$$\underset{x \in C_1, y \in C_2}{Max}\ d(\mathbf{x}, \mathbf{y}) = 2 \cdot \frac{\varepsilon_1}{5}\sqrt{2} < \varepsilon_1.$$

Proof of Lemma 3

$$E[f_n(\mathbf{x})] - E\left[\frac{1}{nh_n^2}\sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right].$$

Note that h_n is a symmetric function in $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Note also that $P\{h_n \in C, \mathbf{x}_i \in B\} = P\{h_n \in C, \mathbf{x}_j \in B\}$ $\forall C, B$ and $\forall i, j \le n$, where C is a borel subset of $\mathscr{R}$ and B is a borel subset of $\mathscr{R}^2$. (It is true because h_n is symmetric in $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ and $\mathbf{x}_i$ and $\mathbf{x}_j$ are identically distributed.)

Thus, $E\left[\frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)\right] = E\left[\frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_j}{h_n}\right)\right]$ $\forall i, j \le n$.

Thus,

$$E[f_n(\mathbf{x})] = E\left[\frac{1}{h_n^2}K\left(\frac{\mathbf{x}-\mathbf{x}_1}{h_n}\right)\right].$$

Construction

Steele[28] proved the following theorem which is given below.

Steele's theorem. Suppose $\mathbf{X}_i$, $1 \le i < \infty$, are independent random variables with distribution μ having compact support in $\mathscr{R}^q$, $q \ge 2$. If the monotone function Ψ satisfies $\Psi(x) \sim x^\alpha$ as $x \to 0$ for some $0 < \alpha < q$, then with probability 1

$$\lim_{n\to\infty} n^{-(q-\alpha)/q} M(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = c(\alpha, q) \int_{\mathscr{R}^q} f(x)^{(q-\alpha)/q}\,dx.$$

Here f denotes the density of the absolutely continuous part of μ and $c(\alpha, q)$ denotes a strictly positive constant which depends only on the power α and the dimension q. $M(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is the length of the minimal spanning tree.

In our case $M(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = l_n$, $q = 2$, $\alpha = 1$. Therefore, $\lim_{n\to\infty} l_n = n^{1/2}k_1$ where k_1 is a positive constant.

Let us define two sequences a_n and c_n for every n such that $a_n = (n^{1/2} - n^{1/3})k_1$ and $c_n = (n^{1/2} - n^{1/3})k_1$. Therefore, for every $\varepsilon > 0$, $\exists M_0 > 0$ such that

$$P(a_n \le l_n \le c_n) \ge 1 - \varepsilon \quad \forall n \ge M_0.$$

Note that $(c_n/a_n) \to 1$ as $n \to \infty$.

Observe that $a_n \to \infty$ and $c_n \to \infty$ as $n \to \infty$. Again $(c_n/n) \to 0$ and $(a_n/n) \to 0$ as $n \to \infty$.

$$\text{Let } \gamma_n = \sqrt{\frac{a_n}{n}} \quad \text{and} \quad t_n = \sqrt{\frac{c_n}{n}}.$$

Now $P(a_n \le l_n \le c_n) = 1 - \varepsilon$ as $n \to \infty$.

$$\text{So } P\left(\sqrt{\frac{a_n}{n}} \le \sqrt{\frac{l_n}{n}} \le \sqrt{\frac{c_n}{n}}\right) = 1 - \varepsilon \text{ as } n \to \infty.$$

$$\text{Thus, } P(t_n \ge h_n \ge \gamma_n) = 1 - \varepsilon \text{ as } n \to \infty.$$

Now $t_n \to 0$ as $n \to \infty$ since $(c_n/n) \to 0$.

Similarly $\gamma_n \to 0$ as $n \to \infty$ since $(a_n/n) \to 0$.

Now $nt_n^2 \to \infty$ and $n\gamma_n^2 \to \infty$ as $n \to \infty$ since $c_n \to \infty$ and $a_n \to \infty$, respectively. Now $(t_n/\gamma_n) \to 1$ as $n \to \infty$ since $(c_n/a_n) \to 1$ as $n \to \infty$.

Proof of Lemma 5

$$\begin{aligned}
\int_{\Gamma_{1n}} g(\mathbf{x}_1, h_n)\,dS_n &= \int_0^{\gamma_n}\int_{\mathscr{R}^2} \frac{1}{\alpha^2}K\left(\frac{\mathbf{x}-\xi}{\alpha}\right)dS_n \\
&= \int_0^{\gamma_n}\int_{\mathscr{R}^2} \frac{1}{\alpha^2}K\left(\frac{\mathbf{x}-\xi}{\alpha}\right)\mu_{1n}[d\xi, \alpha]\,d\beta_n(\alpha) \\
&= \int_0^{\gamma_n} \frac{1}{4\alpha^2}\mu_{1n}[\mathscr{N}_x(\alpha), \alpha]\,d\beta_n(\alpha) \\
&\le \int_0^{\gamma_n} \frac{1}{4\alpha^2}\alpha^2 M_1\,d\beta_n(\alpha) \quad \forall n \ge M_2 = \frac{M_1}{4}P(h_n \le \gamma_n) \\
&\le \frac{M_1}{4}\varepsilon \quad \forall n \ge M_2.
\end{aligned}$$

Similarly it can be shown that:

$$\int_{\Gamma_{2n}} g(\mathbf{x}_1, h_n)\,dS_n \le \frac{M_1}{4}\varepsilon \quad \forall n \ge M_2.$$

Proof of Lemma 6

Now

$$\begin{aligned}
&\int_{\mathscr{R}^2\times A_n} \frac{1}{t_n^2}K\left(\frac{\mathbf{x}-\xi}{\gamma_n}\right)S_n(d\xi, \alpha)) \\
&= \int_{\mathscr{R}^2}\left[\int_{A_n}\frac{1}{t_n^2}K\left(\frac{\mathbf{x}-\xi}{\gamma_n}\right)\nu_{\xi n}(d\alpha)\right]f(\xi)\,d\xi \\
&= \int_{\mathscr{R}^2}\frac{1}{t_n^2}K\left(\frac{\mathbf{x}-\xi}{\gamma_n}\right)\nu_{\xi n}(A_n)f(\xi)\,d\xi = \int_{\mathscr{R}^2} b_{1n}(\xi, \mathbf{x})b_{2n}(\xi)\,d\xi.
\end{aligned}$$

Now

$$\left| \int_{\mathcal{R}^2} b_{1n}(\xi, \mathbf{x}) b_s(\xi)\, d\xi - \int_{\mathcal{R}^2} b_{1s}(\xi, \mathbf{x})\, d\xi \right|$$

$$\leq \int_{\mathcal{R}^2} b_{1s}(\xi, \mathbf{x}) |b_n(\xi) - 1|\, d\xi \leq \int_{\mathcal{R}^2} b_{1n}(\xi, \mathbf{x}) \varepsilon\, d\xi \quad \forall n \geq M_2$$

$$= \varepsilon \int_{\mathcal{R}^2} b_{1n}(\xi, \mathbf{x})\, d\xi \leq \varepsilon(f(\mathbf{x}) - \delta_2) \quad \forall n \geq Max(M_2, M_4).$$

Similarly it can be shown that:

$$\int_{\mathcal{R}^2 \times A_n} \frac{1}{\gamma_n^2} K\left(\frac{\mathbf{x} - \xi}{\tau_n}\right) S_s(d(\xi, z)) \leq \varepsilon(f(\mathbf{x}) + \delta_2)$$

$$\forall n \geq Max(M_3, M_5).$$

About the Author—Professor B.B. CHAUDHURI received his B.Sc. (Hons), B. Tech and M. Tech degrees from Calcutta University, India, in 1969, 1972 and 1974, respectively, and his Ph.D. degree from the Indian Institute of Technology, Kanpur, in 1980. He joined the Indian Statistical Institute in 1978 where he served as the Project Coordinator and Head of the National Nodal Center for Knowledge Based Computing. Currently, he is the head of Computer Vision and Pattern Recognition Unit of the Institute. His research interests include pattern recognition, image processing, computer vision, natural language processing and digital document processing including OCR. He has published 140 research papers in reputed International Journals and has authored a book entitled *Two Tone Image Processing and Recognition* (Wiley Eastern, 1993). He was awarded the *Sir J. C. Bose Memorial Award* for best engineering science oriented paper published in IIETE in 1986 and the *M. N. Saha Memorial Award* (twice) for best application oriented papers published in 1989 and 1991. In 1992 he won the prestigious *Homi Bhabha Fellowship* for working on OCR of the Indian Languages and computer communication for the blind. He has been selected for the *Hari Om Ashram Prerit Dr Vikram Sarabhai Research Award* for the year 1995 for his outstanding achievements in the fields of electronics, information and telematics. As a Leverhulme visiting fellow, he worked at Queen's University, U.K. Also, he worked as a visiting faculty member at GSF, Munich and guest scientist at the Technical University of Hannover during 1986–1988 and again in 1990–1991. He is a Senior member of IEEE, member secretary (Indian Section) of International Academy of Sciences, Fellow of National Academy of Sciences (India) and Fellow of the Indian National Academy of Engineering. He is serving as associate editor of the journals *Pattern Recognition* and *Vivek* as well as guest editor of a special issue of *Journal IETE* on fuzzy systems.

About the Author—Dr D. CHAUDHURI was born in Bolpur (Santiniketan), India. He received a Bachelor of Mathematics (Hons) from Visva-Bharati University, Santiniketa, in 1984 and M.Sc. (Applied Mathematics) from Jadavpur University, Calcutta, in 1987. He worked as a regular research worker in the Department of Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Calcutta, from 1988 to 1995. He received his Ph.D. degree from Jadavpur University, Calcutta, in 1995. Currently he is a project scientist in the ISI-ADRIN Project, Department Space Science, Government of India, at Computer Vision and Pattern Recognition Unit, ISI, Calcutta. His fields of interests are pattern recognition, image processing and computer vision.

About the Author—Dr C. A. MURTHY was born in Ongole, India, in 1958. He obtained the B. Stat (Hons.), M. Stat and Ph.D. degrees from Indian Statistical Institute, Calcutta, India, in 1979, 1980 and 1989, respectively. He visited the Michigan State University, East Lansing, U.S.A., as UNDP fellow in 1991–1992. Currently, he is an associate professor in the Machine Intelligence Unit of the Indian Statistical Institute, Calcutta, India. His fields of interests include pattern recognition, image processing, fuzzy sets, neural networks, genetic algorithms and fractals.