On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions

ROBERT P. W. DUIN

Abstract—**Parzen estimators are often used for nonparametric estimation of probability density functions. The smoothness of such an estimation is controlled by the smoothing parameter. A problem-dependent criterion for its value is proposed and illustrated by some examples. Especially in multimodal situations, this criterion led to good results.**

Manuscript received November 16, 1973; revised November 20, 1975.

The author is with the Department of Applied Physics, Delft University of Technology, Delft, The Netherlands.

Index Terms—**Data analysis, multidimensional probability density function, multimodal probability density function, Parzen estimators, pattern recognition, smoothing.**

INTRODUCTION

An important problem in the field of pattern recognition is the estimation of the probability density function $F(x)$ from a number of randomly selected samples $x_1,x_2,\cdots,x_n$. If little *a priori* knowledge about $F(x)$ is available, a nonparametric estimate of $F(x)$ can be useful. A well-known method is the Parzen estimation:

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} h(n)^{-m} K\left(\frac{||x - x_i||}{h(n)}\right) \tag{1}$$

in which K, the kernel, is an arbitrary bounded probability density, n is the number of samples, $h(n)$ is a sequence of positive numbers, and $||\cdot||$ is the Euclidian norm. It has been proved by Parzen [5] for the one-dimensional case and extended by Murthy [4] and others for the multidimensional case that

$$\lim_{n\to\infty} h(n) = 0 \tag{1a}$$

guarantees asymptotic unbiasedness,

$$\lim_{n\to\infty} nh^m(n) = \infty \tag{1b}$$

in which m is the dimensionality guarantees asymptotic consistency, and

$$\lim_{n\to\infty} nh^{2m}(n) = \infty \tag{1c}$$

guarantees uniform consistency.

Examples for K are $\exp(-||x||)$, the normal density, and $\sin^2(||x||)/||x||^2$. As can be understood from (1), the degree of smoothing of the estimate $\hat{F}(x)$ is controlled by $h(n)$. For that reason, h is called the smoothing parameter. It is desirable that the data play a role in the amount of smoothing, so h should not only be a function of n, but also of $x_1,x_2,\cdots,x_n$. A method by Loftsgaarden and Quesenberry [3] allows $h(n)$ to depend on the data, but also on the point of estimation x. For applications in which a rather simple analytical expression for $\hat{F}(x)$ is needed, this is an undesirable situation because, for each point of estimation, first the smoothing parameter has to be calculated. Koontz and Fukunaga [2] used a method in which h is a function of n and of the estimated covariance matrix $\hat{C}$:

$$h(n) = \left(\frac{1}{m}\operatorname{tr}\hat{C}\right)^{1/2} n^{-\alpha/m}. \tag{2}$$

α is a constant that satisfies $0 < \alpha < 0.5$. The method is optimized for an underlying distribution which is normal. In nonnormal, e.g., multimodal, situations, this method can yield values of h which lead to poor estimates of the probability density function. This will be shown later by some examples.

METHOD

We introduce a method which is more problem-dependent than the method given by Fukunaga and Koontz.

We take the product of the estimated densities in the sample points and try to find that value of h which optimizes this product:

$$\max_h L'(h) = \prod_{j=1}^{n} \hat{F}(x_j).$$

So the optimal value of h is, in this sense, not only a function of n, but also of the data. Substituting (1) for $\hat{F}(x)$ and using the normal density as a kernel gives

$$\max_h L'(h) = \prod_{j=1}^{n} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(h\sqrt{2\pi})^m} \exp\left\{-\frac{||x_j - x_i||}{2h^2}\right\}. \quad (3)$$

It can be easily verified that

$$\lim_{h \to 0} L'(h) = \infty$$

and $L'(h)$ is finite for $h \neq 0$, so $L'(h)$ has an absolute maximum for $h = 0$. This corresponds (see (1) in which K is the normal density) with an estimate of delta pulses at the sample points and a zero estimate elsewhere. This is caused by the fact that each term in the product of (3) becomes infinite if h becomes zero because in each term $||x_j - x_i||$ becomes zero of $i = j$. In order to avoid this undesirable situation, we will omit the contribution of the sample itself in the estimation of the density at that point:

$$\max_h L(h) = \prod_{j=1}^{n} \hat{F}_j(x_j) \quad (4)$$

in which

$$\hat{F}_j(x) = \sum_{i \neq j}^{n} \frac{1}{(h\sqrt{2\pi})^m} \exp\left\{-\frac{||x - x_i||^2}{2h^2}\right\}.$$

From a number of experiments, this criterion appeared to be useful. The product in (4) guarantees a nonzero density for each x_j. A disadvantage is, therefore, that one "wild shot" considerably influences $L(h)$ and thus the optimal value of h. However, an alternative without this effect, e.g., the replacement of the product by a summation, will have other, much larger drawbacks. In that case, a number of zero or close to zero densities are allowed. $L(h)$ will mainly be influenced by points on small distances of each other because in these points the density can grow most easily. This causes h to be smaller than by using criterion (4) and leads to worse estimates.

Examples of this are given in [8].

We will now present some examples of the use of (4) and will compare the results with those obtained by using (2).

A. Two Cluster Case

Suppose there are two clusters, each consisting of $n (n > 1)$ sample points. The sample distance within a cluster is zero. The distance between the clusters is a.

$$L(h) = \prod_{i=1}^{2n} \left\{\frac{n-1}{(h\sqrt{2\pi})^m} \exp(0) + \frac{n}{(h\sqrt{2\pi})^m} \exp(-a^2/2h^2)\right\}.$$

For small values of h,

$$L(h) = \prod_{i=1}^{2n} \left(\frac{n-1}{h\sqrt{2\pi}}\right) = \left(\frac{n-1}{h\sqrt{2\pi}}\right)^{2n}.$$

This is maximum for $h = 0$. Because this maximum is infinite and $L(h)$ is finite for all values of $h \neq 0$, this maximum is absolute. The estimation in this case is exact: delta pulses in the cluster points.

The method of Koontz and Fukunaga gives

$$h^2 = \frac{a^2}{m} (2n)^{-\alpha/m}$$

which only approaches zero for $m \to \infty$ or $n \to \infty$.

B. Number of Identical Clusters

If the sizes of the clusters are small compared with the between-cluster distances, then, for small values of h, the estimates are only built up by contributions of points belonging to the same cluster. So if $C_1, C_2, \cdots, C_k$ indicate the cluster associations of the various sample point indices,

$$L(h) = \prod_{i \in C_1} \hat{F}_i(x_i) \prod_{i \in C_2} \hat{F}_i(x_i) \cdots \prod_{i \in C_k} \hat{F}_i(x_i)$$

in which

$$\hat{F}_i(x) = \frac{1}{n} \sum_{\substack{j \in C_1 \\ j \neq i}} \frac{1}{h} K\left(\frac{||x - x_j||}{h}\right)$$

if x_i belongs to cluster 1. The number of samples in each cluster is n. If the clusters are identical, $L(h)$ can be approximated by

$$L(h) = \left\{\prod_{i \in C_1} \hat{F}_i(x_i)\right\}^k.$$

This is maximum if $\Pi_{i \in C_1} \hat{F}_i(x_i)$ is maximum. So we find the same solution for h, for small values of h, if we try to estimate the density function of a single cluster or if we do that for a group of identical clusters. The sensitivity of (4) for multimodal situations, of which the above is an example, is in our opinion the main advantage of (4) compared to (2).

C. Normal Distribution

Suppose the underlying distribution $F(x)$ is normal with the unity matrix as covariance matrix. We will experimentally investigate the dependence of the error upon the number of samples.

The error, defined as the nonoverlap of the estimated distribution $\hat{F}(x)$ and the real distribution $F(x)$, can be expressed by

$$D = 1 - \int_{\forall x} \min(F(x), \hat{F}(x)) dx.$$

This can be estimated by generating points x according to $F(x)$ and counting the number of times $F(x)$ is smaller than $\hat{F}(x)$, and generating points according to $\hat{F}(x)$ and counting the number of times $\hat{F}(x)$ is smaller than $F(x)$. In [8] it is shown that if from both distributions q points are generated, the variance in the estimate of D is bounded by

$$\text{var}(\hat{D}) \leq \frac{1}{2q}(1 - D^2).$$

In all our experiments we used $q = 100$.

We will use three estimation methods for $\hat{F}(x)$.

1) Parametric estimation, by assuming that the underlying distribution is normal (which it is indeed) and estimating mean and covariance matrix from the sample set using ML estimators.

2) Nonparametric estimation, using Parzen estimators, with the normal density as a kernel and with (2) as an estimate for h.

3) Same as 2), but with (4) as a criterion for h.

Fig. 1 shows, for the one-dimensional case, how the D behaves as a function of the number of samples. The parametric estimation converges faster. There is no significant difference between the two versions of the Parzen estimation. These experiments and all the following ones are averaged over ten runs in order to reduce the influence of a single extreme result. In Fig. 2 the results of a ten-dimensional experiment are shown. Here it appears that for small sample sets, the nonparametric estimate can be better than the parametric, although the latter made use of the normality. Probably this is caused by the fact that the normal distribution is used as the kernel of the Parzen estimation. As can be expected, the parametric estimation always converges faster.

D. Multimodal Case

In the multimodal case, criterion (4) shows its real power. Again we will illustrate this with a simulation. Our distribution was built up with five m-dimensional normal distributions, each with the unity matrix as a covariance matrix. Their means were located

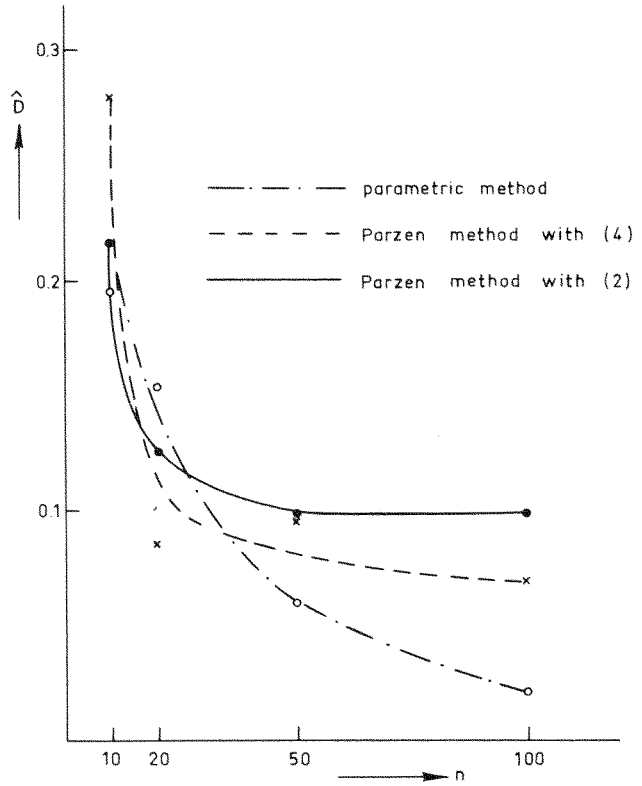

Fig. 1. $\hat{D}$ as a function of the number of learning samples for a one-dimensional normal distribution. The measurement points are averages after 10 runs. There is no significant difference between estimation (2) and criterion (4). The parametric estimation converges significantly faster than the nonparametric ones.

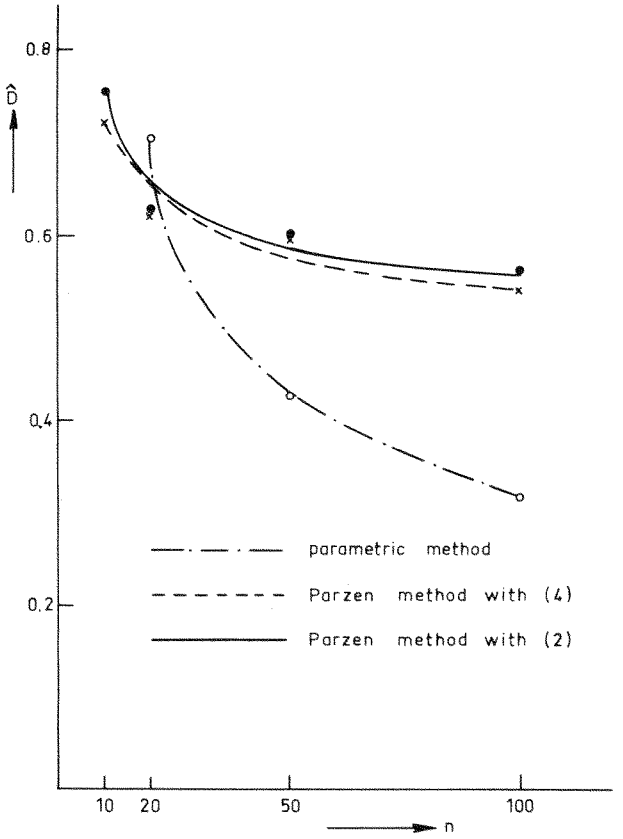

Fig. 2. $\hat{D}$ as a function of the number of learning samples for a ten-dimensional normal distribution. The measurement points are averages after 10 runs. There is no significant difference between estimation (2) and criterion (4). The parametric estimation converges significantly faster than the nonparametric ones.

on a straight line at distances of a. We selected at random n samples from this distribution; (2) and (4) were used for estimation of the smoothing parameter. The estimated values of D for $a = 10$ as a function of n is given in Figs. 3 and 4 for $m = 1$ and $m = 5$. D as a function of a for the one-dimensional case ($m = 1$) and $n = 50$ is given in Fig. 5. All experiments were averaged over ten runs. As can be seen, the asymptotic properties of the proposed criterion (4) are better than those of the method of Koontz and Fukunaga which increases for larger a. This can be understood by considering (2) and (4). Because (2) only depends on the covariance matrix and is not sensitive for multimodality, the estimation of well-separated clusters is poor. Criterion (4), on the other hand, produces in multimodal situations reasonable estimates. An example is given in Fig. 6.

Computational Aspects

The optimization of (4) is always executed by finding the zero crossing(s) of its first derivative, using a special adapted version of the *regula falsi*. In all investigated experiments, only one zero crossing was present, but this is not necessarily true for all problems. In most experiments, only 5–20 iterations were needed to reach an accuracy of 10^{-3} in the value of the smoothing parameter. The computing time using an IBM 360/65 was, for 100 samples, about 0.9 s/iteration. It is, however, necessary to have the complete intersample distance matrix available. This limits the applicability of the method to a few hundred samples.

If one wants to use larger samples sets, it is necessary to divide the sample space in some areas and estimate the probability density function of each area separately.

Applications

The Parzen estimation using (4) as a criterion for the smoothing parameter is well suited for the estimation of multimodal density functions. This can be applied in cluster analysis where one tries to split up the sample set in homogeneous subsets.

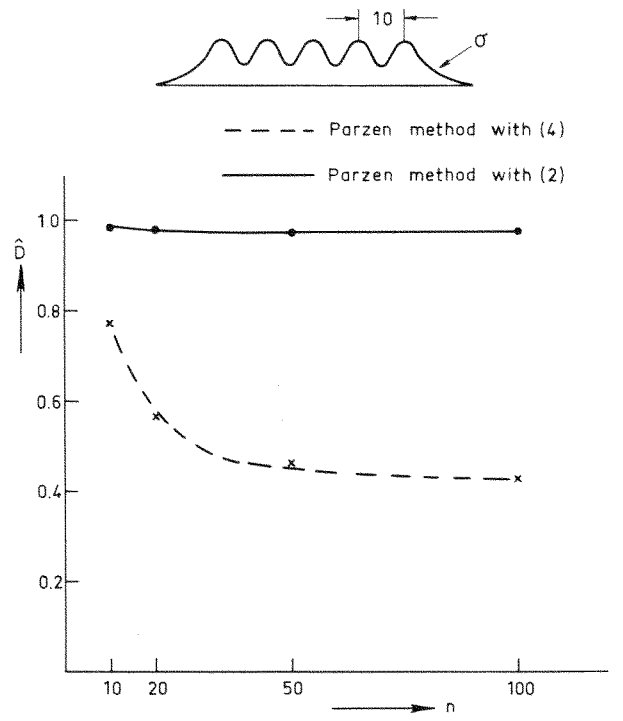

Fig. 3. $\hat{D}$ as a function of the number of learning samples for a one-dimensional distribution consisting of five normal distributions, each with variance one, on a straight line with equal distances of $a = 10$. The measurement points are averages after 10 runs. The Parzen estimation using criterion (4) converges significantly faster than using estimation (2).

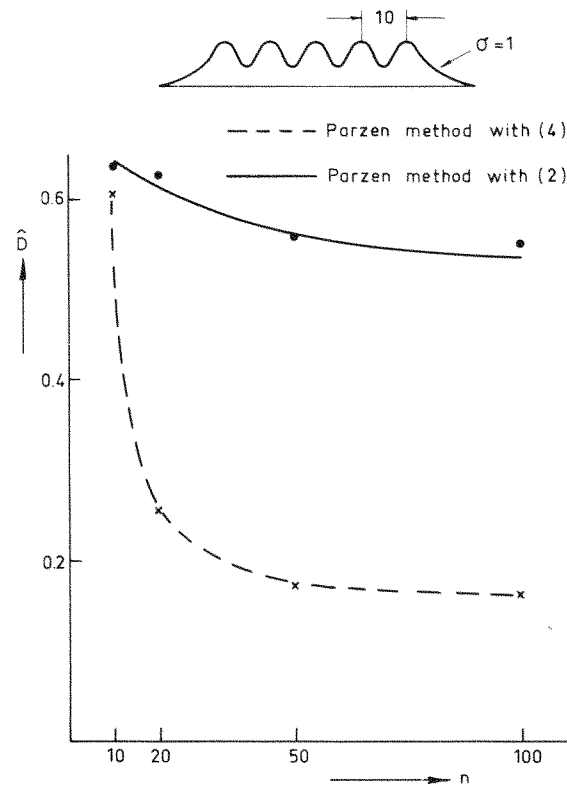

Fig. 4. $\hat{D}$ as a function of the number of learning samples for a five-dimensional distribution consisting of five normal distributions, each with covariance matrix I, on a straight line with equal distances of $a = 10$. The measurement points are the averages after 10 runs. The Parzen estimation using criterion (4) converges significantly faster than using estimation (2).

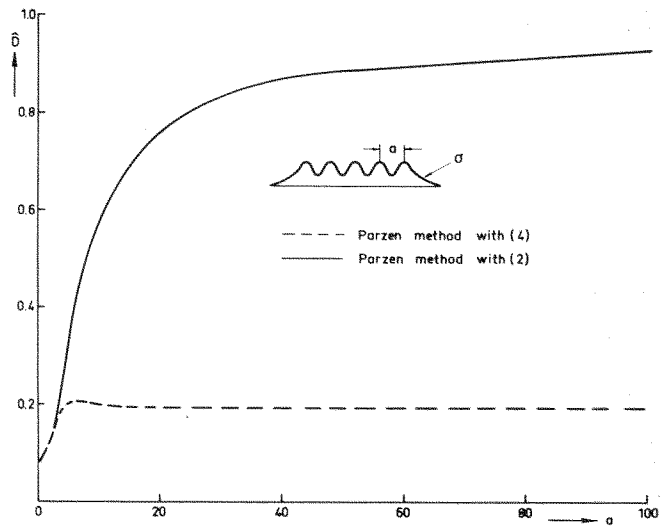

Fig. 5. $\hat{D}$ in the one-dimensional case as a function of a for estimation (2) and criterion (4) for 50 learning samples.

Using a method like that of Gitman and Levine [1], a rather fast cluster analysis can be achieved using the intersample distance matrix and the estimated densities in the sample points. Also, in the field of feature extraction, applications are possible. For instance, Parzen estimation may be used to optimize the distance between the probability density functions of two classes; see Patrick and Fischer [6].

For the separation of classes, Parzen estimations are also used (Specht [7]). Optimization of the smoothing parameter is often obtained by optimizing the classification result of a learning set. The proposed method is at least faster; whether or not this leads to better results on test sets remains to be investigated.

Future Research

The following points are open for further research.

1) Possibilities of estimating different values of the smoothing parameter for each dimension.

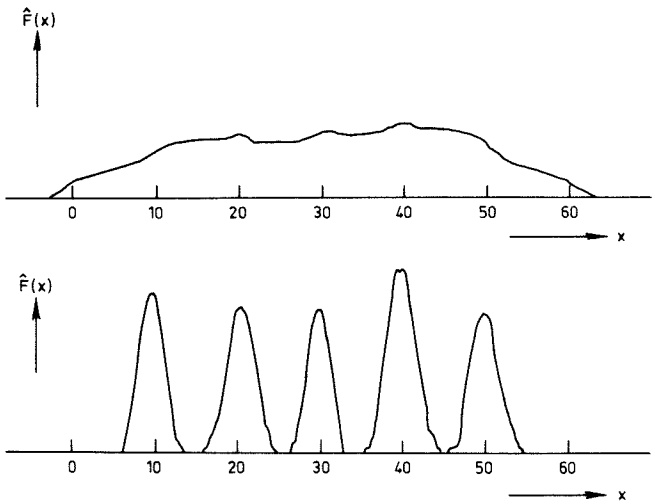

Fig. 6. An example of the density estimate in the one-dimensional case, using estimation (2) (above) and criterion (4) (below) with 100 samples.

2) Computational aspects of using large sample sets as to estimating the smoothing parameter and as to computing the density of a new point.

3) Simplifications of the estimated density function by using a small number of normal distributions.

4) Theoretical investigations of the covergence properties. It would be valuable to prove whether or not the properties of h as a function of the sample size n, given in the Introduction, are true.

REFERENCES

[1] I. Gitman and M. D. Levine, "An algorithm for detecting unimodel fuzzy sets and its application as a clustering technique," *IEEE Trans. Comput.*, vol. C-19, pp. 583–593, July 1970.

[2] W. L. G. Koontz and K. Fukunaga, "Asymptotic analysis of a nonparametric clustering technique," *IEEE Trans. Comput.*, vol. C-21, pp. 967–974, Sept. 1972.

[3] D. O. Loftsgaarden and C. D. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, pp. 1049–1051, 1965.

[4] V. K. Murthy, "Nonparametric estimation of multivariate densities with applications," in *Multivariate Analysis*, P. R. Krishnaiah, Ed. New York: Academic, 1966, pp. 43–48.

[5] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1076, 1962.

[6] E. A. Patrick and F. P. Fischer, "Nonparametric feature selection," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 577–584, Sept. 1969.

[7] D. F. Specht, "Generation of polynomical discriminant functions for pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-16, pp. 308–319, Apr. 1967.

[8] R. P. W. Duin, "A criterion for the smoothing parameter for Parzen estimators of probability density functions," Pattern Recognition Group, Dep. Appl. Phys., Delft Univ. Technol., Delft, The Netherlands, Internal Rep., Sept. 1975.