The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition

KEINOSUKE FUKUNAGA, SENIOR MEMBER, IEEE, AND LARRY D. HOSTETLER, MEMBER, IEEE

Abstract—**Nonparametric density gradient estimation using a generalized kernel approach is investigated. Conditions on the kernel functions are derived to guarantee asymptotic unbiasedness, consistency, and uniform consistency of the estimates. The results are generalized to obtain a simple mean-shift estimate that can be extended in a *k*-nearest-neighbor approach. Applications of gradient estimation to pattern recognition are presented using clustering and intrinsic dimensionality problems, with the ultimate goal of providing further understanding of these problems in terms of density gradients.**

I. Introduction

NONPARAMETRIC estimation of probability density functions is based on the concept that the value of a density function at a continuity point can be estimated using the sample observations that fall within a small region around that point. This idea was perhaps first used in Fix and Hodges' [1] original work. Rosenblatt [2], Parzen [3], and Cacoullos [4] generalized these results and developed the Parzen kernel class of density estimates. These estimates were shown to be asymptotically unbiased, consistent in a mean-square sense, and uniformly consistent (in probability). Additional work by Nadaraya [5], later extended by Van Ryzin [6] to n dimensions, showed strong and uniformly strong consistency (with probability one).

Similarly, the gradient of a probability density function can be estimated using the sample observations within a small region. In this paper we will use these concepts and results to derive a general kernel class of density gradient estimates. Although Bhattacharyya [7] and Schuster [8] considered estimating the density and all its derivatives in the univariate case, we are interested in the multivariate gradient, and our results follow more closely the Parzen kernel estimates previously mentioned.

In Section II, we will develop the general form of the kernel gradient estimates and derive conditions on the kernel functions to assure asymptotically unbiased, consistent, and uniformly consistent estimates. By examining the form of the gradient estimate when a Gaussian kernel is used, in Section III we will be able to develop a simple mean-shift class of estimates. The approach uses the fact that the expected value of the observations within a small region about a point can be related to the density gradient at that point. A simple modification results in a k-nearest-neighbor mean-shift class of estimates. This modification is similar to Loftsgaarden and Quesenberry's [9] k-nearest-neighbor density estimates, in that the number of observations within the region is fixed whereas in the previous estimates the volume of the region was fixed.

In Section IV, applications of density gradient estimation to pattern recognition problems are investigated. By taking a mode-seeking approach, a recursive clustering algorithm is developed, and its properties studied. Interpretations of our results are presented along with examples. Applications to data filtering and intrinsic dimensionality determination are also discussed. Section V is a summary.

II. Estimation of the Density Gradient

In most pattern recognition problems, very little if any information is available as to the true probability density function or even as to its form. Due to this lack of knowledge about the density, we have to rely on nonparametric techniques to obtain density gradient estimates. A straightforward approach for estimating a density gradient would be to first obtain a differentiable nonparametric estimate of the probability density function and then take its gradient. The nonparametric density estimates we will use are Cacoullos' [4] multivariate extensions of Parzen's [3] univariate kernel estimates.

A. Proposed Gradient Estimates

Let $X_1,X_2,\cdots,X_N$ be a set of N independent and identically distributed n-dimensional random vectors. Cacoullos [4] investigated multivariate kernel density estimators of the form

$$\hat{p}_N(X) \equiv (Nh^n)^{-1} \sum_{j=1}^{N} k(h^{-1}(X - X_j)) \tag{1}$$

where $k(X)$ is a scalar function satisfying

$$\sup_{Y \in R^n} |k(Y)| < \infty \tag{2}$$

$$\int_{R^n} |k(Y)|\, dY < \infty \tag{3}$$

$$\lim_{\|Y\| \to \infty} \|Y\|^n k(Y) = 0 \tag{4}$$

$$\int_{R^n} k(Y)\, dY = 1 \tag{5}$$

and where $\|\cdot\|$ is the ordinary Euclidean norm, and R^n is the n-dimensional feature space. The parameter h, which

Manuscript received February 1, 1973; revised June 13, 1974. This work was supported in part by the National Science Foundation under Grant GJ-35722 and in part by Bell Laboratories, Madison, N.J.

K. Fukunaga is with the School of Electrical Engineering, Purdue University, Lafayette, Ind. 47907.

L. D. Hostetler was with Bell Laboratories, Whippany and Madison, N.J. He is now with the Sandia Laboratories, Albuquerque, N. Mex. 87115.

is a function of the sample size N, is chosen to satisfy

$$\lim_{N\to\infty} h(N) = 0 \tag{6}$$

to guarantee asymptotic unbiasedness of the estimate. Mean-square consistency of the estimate is assured by the condition

$$\lim_{N\to\infty} Nh^n(N) = \infty. \tag{7}$$

Uniform consistency (in probability) is assured by the condition

$$\lim_{N\to\infty} Nh^{2n}(N) = \infty, \tag{8}$$

provided the true density $p(X)$ is uniformly continuous. Additional conditions were found by Van Ryzin [6] to assure strong and uniformly strong consistency (with probability one).

Motivated by this general class of nonparametric density estimates, we use the differentiable kernel function and then estimate the density gradient as the gradient of the density estimate (1). This gives the density gradient estimate

$$\hat{\nabla}_x p_N(X) \equiv (Nh^n)^{-1} \sum_{j=1}^{N} \nabla_x k(h^{-1}(X - X_j)) \tag{9}$$

$$= (Nh^{n+1})^{-1} \sum_{j=1}^{N} \nabla k(h^{-1}(X - X_j)) \tag{10}$$

where

$$\nabla k(Y) \equiv \left(\frac{\partial k(Y)}{\partial y_1}, \frac{\partial k(Y)}{\partial y_2}, \cdots, \frac{\partial k(Y)}{\partial y_n}\right)^T \tag{11}$$

and ∇_x is the usual gradient operator with respect to $x_1, x_2, \cdots, x_n$.

Equation (10) is the general form of our density gradient estimates. As in density estimation, this is a kernel class of estimates, and various kernel functions $k(Y)$ may be used. Conditions on the kernel functions and $h(N)$ will now be derived to guarantee asymptotic unbiasedness, consistency, and uniform consistency of the estimate.

B. Asymptotic Unbiasedness

The basic result needed for the proof of asymptotic unbiasedness is [4, theorem 2.1]. This theorem, given here without proof, states that if the function $k(X)$ satisfies conditions (2)–(4), $h(N)$ satisfies condition (6), and $g(X)$ is any other function such that

$$\int_{R^n} |g(Y)|\, dY < \infty, \tag{12}$$

then the sequence of functions $g_N(X)$ defined by

$$g_N(X) \equiv h^{-n}(N) \int_{R^n} k(h^{-1}(N)Y) g(X - Y)\, dY \tag{13}$$

converges at every point of continuity of $g(X)$ to

$$\lim_{N\to\infty} g_N(X) = g(X) \int_{R^n} k(Y)\, dY. \tag{14}$$

Using this we will be able to show that the estimate $\hat{\nabla}_x p_N(X)$ is asymptotically unbiased for a class of density functions that satisfy

$$\lim_{\|X\|\to\infty} p(X) = 0. \tag{15}$$

Although the condition (15) must be imposed on the density function, practically all density functions satisfy this condition. Thus

$$\lim_{N\to\infty} E\{\hat{\nabla}_x p_N(X)\} = \nabla_x p(X) \tag{16}$$

at the points of continuity of $\nabla_x p(X)$, provided that 1) $p(X)$ satisfies (15); 2) $h(N)$ goes to zero as $N \to \infty$ (see (6)); 3) $k(X)$ satisfies (2)–(5); and 4) $\int_{R^n} |\partial p(Y)/\partial y_i|\, dY < \infty$, for $i = 1,2,\cdots,n$.

The proof is given in the Appendix.

C. Consistency

The estimate $\hat{\nabla}_x p_N(X)$ is consistent in quadratic mean,

$$\lim_{N\to\infty} E\{\|\hat{\nabla}_x p_N(X) - \nabla_x p(X)\|^2\} = 0 \tag{17}$$

at the points of continuity of $\nabla_x p(X)$, provided that the following conditions are satisfied

1)
$$\lim_{N\to\infty} h(N) = 0 \tag{18}$$

$$\lim_{N\to\infty} Nh^{n+2}(N) = \infty \tag{19}$$

and in addition to (2)–(5) the kernel function is such that

2)
$$\sup_{Y\in R^n} |k_i'(Y)| < \infty \tag{20}$$

$$\int_{R^n} |k_i'(Y)|\, dY < \infty \tag{21}$$

$$\lim_{\|Y\|\to\infty} \|Y\|^n k_i'(Y) = 0 \tag{22}$$

where

$$k_i'(Y) \equiv \frac{\partial k(Y)}{\partial y_i}. \tag{23}$$

The proof is given in the Appendix.

The additional power of two in the condition (19) on $h(N)$ comes from estimating gradients and not just the density so that in (17),

$$\left[\frac{\partial}{\partial x_i} k(h^{-1}(X - Y))\right]^2 = h^{-2}[k_i'(h^{-1}(X - Y))]^2 \tag{24}$$

with the additional h^{-2} being generated. This additional power requires $h(N)$ to go to zero slightly slower than in the previous case (7) for density estimation alone.

Since we may want to use the estimate of the density gradient over the entire space in the same application and not just be satisfied with pointwise properties, we would like to determine the conditions under which the gradient estimate is uniformly consistent.

D. Uniform Consistency

The gradient estimate $\hat{\nabla}_x p_N(X)$ is uniformly consistent. That is, for every $\varepsilon > 0$,

$$\lim_{N\to\infty} \Pr\{\sup_{R^n} \|\hat{\nabla}_x p_N(X) - \nabla_x p(X)\| > \varepsilon\} = 0. \tag{25}$$

The conditions to be satisfied are listed as follows

1)
$$\lim_{N\to\infty} h(N) = 0 \tag{26}$$

$$\lim_{N\to\infty} Nh^{2n+2}(N) = \infty \tag{27}$$

2) the characteristic function

$$G(W) = \int_{R^n} \exp(jW^TX)\nabla_x k(X)\,dX$$

of $\nabla_x k(X)$ is absolutely integrable (i.e.,

$$\int_{R^n} \|G(W)\|\,dW < \infty, \qquad \text{where } W = [\omega_1 \cdots \omega_n]^T) \tag{28}$$

3)
$$\nabla_x p(X) \text{ is uniformly continuous.} \tag{29}$$

The proof follows from the definition of $\hat{\nabla}_x p_N(X)$ and the properties of characteristic functions, the details of which are given in the Appendix. Again we see that the additional power of two in the condition (27) on $h(N)$ is present.

III. Special Kernel Functions

Having derived a general class of probability density gradient estimates and shown it to have certain desirable properties, we will now investigate specific kernel functions in order to better understand the underlying process involved in gradient estimation. Ultimately this results in a simple and very intuitive mean-shift estimate for the density gradient. A simple modification then extends this in a k-nearest-neighbor approach to gradient estimation much in the same manner as Loftsgaarden and Quesenberry's [9] extension of kernel density estimates.

A. Mean-Shift Gradient Estimates

The Gaussian kernel function is perhaps the best known differentiable multivariate kernel function satisfying the conditions for asymptotic unbiasedness, consistency, and uniform consistency of the density gradient estimate. The Gaussian probability density kernel function with zero mean and identity covariance matrix is

$$k(X) \equiv (2\pi)^{-n/2} \exp(-\tfrac{1}{2}X^TX). \tag{30}$$

Taking the gradient of (30) and substituting it into (10) for $\nabla k(X)$, we obtain as the estimate of the density gradient

$$\hat{\nabla}_x p_N(X) = N^{-1} \sum_{i=1}^{N} (X_i - X)(2\pi)^{-n/2} h^{-(n+2)} \cdot \exp\left[-(X - X_i)^T\left(\frac{X - X_i}{2h^2}\right)\right]. \tag{31}$$

An intuitive interpretation of (31) is that it is essentially a weighted measure of the mean shift of the observations about the point X. The shift $X_i - X$ of each point from X is calculated and multiplied by the weighting factor $h^{-(n+2)} \cdot (2\pi)^{-n/2} \exp(-(X - X_i)^T(X - X_i)/2h^2)$. The sample mean of these weighted shifts is then taken as the gradient estimate.

The same general form will result if the kernel function is of the form

$$k(X) = g(X^TX). \tag{32}$$

The most simple kernel with this form is

$$k(X) = \begin{cases} c(1 - X^TX), & X^TX \le 1 \\ 0, & X^TX > 1 \end{cases} \tag{33}$$

where

$$c = \pi^{-n/2}\left(\frac{n+2}{2}\right)\Gamma\left(\frac{n+2}{2}\right) \tag{34}$$

is the normalizing constant required to make the kernel function integrate to one and $\Gamma(\cdot)$ is the gamma function. This kernel function can easily be shown to satisfy all the conditions for asymptotic unbiasedness, consistency, and uniform consistency of the gradient estimate and is, in a sense, quite similar to the asymptotic optimum product kernel discussed by Epanechnikov [10].

Substituting (33) into (10), we obtain as the gradient estimate

$$\begin{aligned}\hat{\nabla}_x p_N(X) &= (Nh^{n+2})^{-1} 2c \sum_{X_i \in S_h(X)} (X_i - X) \\ &= \left(\frac{k}{Nv_h(X)}\right)\frac{n+2}{h^2}\left(\frac{1}{k}\sum_{X_i \in S_h(X)} (X_i - X)\right)\end{aligned} \tag{35}$$

where

$$v_h(X) \equiv \int_{S_h(X)} dY = \frac{h^n \pi^{n/2}}{\Gamma(n + 2/2)} \tag{36}$$

is the volume of the region

$$S_h(X) \equiv \{Y : (Y - X)^T(Y - X) \le h^2\} \tag{37}$$

and k is the number of observations falling within $S_h(X)$ and, therefore, the number in the sum.

Equation (35) provides us with an excellent interpretation of the gradient estimation process. The last term in (35) is the sample mean shift

$$M_h(X) \equiv \frac{1}{k}\sum_{X_i \in S_h(X)} (X_i - X) \tag{38}$$

of the observations in the small region $S_h(X)$ about X. Clearly, if the gradient or slope is zero, corresponding to a uniform density over the region $S_h(X)$, the average mean shift would be zero due to the symmetry of the observations near X. However, with a nonzero density gradient pointing in the direction of most rapid increase of the probability density function, on the average more observations should fall along its direction than elsewhere in $S_h(X)$. Correspondingly, the average mean shift should point in that direction and have a length proportional to the magnitude of the gradient.

B. Mean-Shift Normalized Gradient Estimates

Examining the proportionality constant in (35), we see that it contains a term identical to the probability density estimate using a uniform kernel function over the region

$S_h(X)$,

$$\hat{p}_N(X) = \frac{k}{Nv_h(X)}. \tag{39}$$

By taking this to the left side of (35) and using the properties of the function ln y, we see that the mean shift can be used as an estimate of the normalized gradient

$$\frac{\nabla_x p(X)}{p(X)} = \nabla_x \ln p(X) \tag{40}$$

$$\hat{\nabla}_x \ln p_N(X) \equiv \frac{n+2}{h^2} M_h(X). \tag{41}$$

This mean-shift estimate of the normalized gradient has a pleasingly simple and easily calculated expression (41). It can also be given the same intuitive interpretation that was just given in the previous section for the gradient estimate. The normalized gradient can be used in most applications in place of the regular gradient, and, as will be seen in Section IV, it has desirable properties for pattern recognition applications.

This mean-shift estimate can easily be generalized to a k-nearest-neighbor approach to normalized gradient estimation. Letting h be replaced by the value of $d_k(X)$, the distance to the k-nearest-neighbor of X, and $S_h(X)$ be replaced by

$$S_{d_k}(X) \equiv \{Y : \|Y - X\| \leq d_k\} \tag{42}$$

we obtain for the k-nearest-neighbor mean-shift estimate of $\nabla_x \ln p(X)$,

$$\hat{\nabla}_x \ln p_N(X) \equiv \frac{n+2}{d_k^2} \frac{1}{k} \sum_{X_i \in S_{d_k}(X)} (X_i - X). \tag{43}$$

Thus we have developed both a kernel and a k-nearest-neighbor approach to normalized gradient estimation.

What, if any, limiting properties, such as asymptotic unbiasedness, consistency, and uniform consistency, carry over from the kernel estimates to the k-nearest-neighbor case is still an open research problem.

IV. Applications

In this section we will show how gradient estimates can be applied to pattern recognition problems.

A. A Gradient Clustering Algorithm

From Fig. 1, we see that one method of clustering a set of observations into different classes would be to assign each observation to the nearest mode along the direction of the gradient at the observation points. To accomplish this, one could move each observation a small step in the direction of the gradient and iteratively repeat the process on the transformed observations until tight clusters result near the modes. Another approach would be to shift each observation by some amount proportional to the gradient at the observation point. This transformation approach is the one we will investigate in this section since it is intuitively appealing and will be shown to have good physical motivation behind its application.

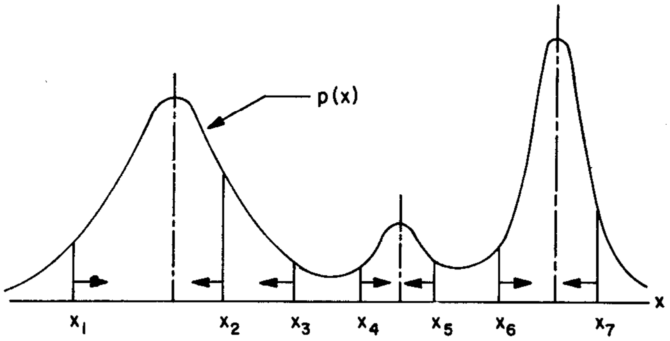

Fig. 1. Gradient mode clustering.

Letting

$$X_j^0 \equiv X_j, \qquad j = 1,2,\cdots,N \tag{44}$$

we will transform each observation recursively according to the clustering algorithm

$$X_j^{i+1} = X_j^i + a\nabla_x \ln p(X_j^i). \tag{45}$$

where a is an appropriately chosen positive constant to guarantee convergence.

This is the n-dimensional analog of the linear iteration technique for stepping into the roots of the equation

$$\nabla_x p(X) = 0 \tag{46}$$

and equivalently the modes of the mixture density $p(X)$. Although local minima are also roots of (46), it is easy to see that the algorithm (45) will move observations away from these points since the gradients point away from them. Thus after each iteration each observation will have moved closer to its parent mode or cluster center.

The use of the normalized gradient $\nabla_x \ln p(X) = \nabla_x p(X)/p(X)$ in place of $\nabla_x p(X)$ in (45) can be justified from the following four points of view.

1) The first is that in the tails of the density and near local minima where $p(X)$ is relatively small it will be true that $\nabla_x \ln p(X) = \nabla_x p(X)/p(X) > \nabla_x p(X)$. Thus the corresponding step size for the same gradient will be greater than that near a mode. This will allow observations far from the mode or near a local minimum to move towards the mode faster than using $\nabla_x p(X)$ alone.

2) The second reason for using the normalized gradient can be seen if a Gaussian density with mean M and identity covariance matrix is considered

$$p(X) = (2\pi)^{-n/2} \exp\left[-\tfrac{1}{2}(X - M)^T(X - M)\right]. \tag{47}$$

Then if a is taken to be one, (45) becomes

$$X_j^1 = X_j + \nabla_x \ln p(X_j) = X_j - (X_j - M) = M. \tag{48}$$

Thus for this Gaussian density the correct choice of a allows the clusters to condense to single points after one iteration. This fact gives support for its use in the general mixture density problem in which the individual mixture component densities often look Gaussian.

3) The third reason for using the normalized gradient is that we can show convergence in a mean-square sense. In other words, if we make the transformation of random variables

$$Y = X + a\nabla_x \ln p(X) \tag{49}$$

then by selecting a proper a the covariance of our samples will get smaller,

$$E\{Y - M_Y)^T(Y - M_Y)\} \leq E\{(X - M_X)^T(X - M_X)\} \tag{50}$$

representing a tightening up of the clusters, where

$$M_Y \equiv E\{Y\} \text{ and } M_X \equiv E\{X\}. \tag{51}$$

Letting

$$Z \equiv \nabla_x \ln p(X) \tag{52}$$

and

$$M_Z \equiv E\{Z\} \tag{53}$$

we obtain from (49),

$$\begin{aligned} E\{(Y - M_Y)^T(Y - M_Y)\} &= E\{(X - M_X)^T(X - M_X)\} \\ &\quad + 2aE\{(X - M_X)^T(Z - M_Z)\} \\ &\quad + a^2E\{(Z - M_Z)^T(Z - M_Z)\}. \end{aligned} \tag{54}$$

Now the ith component of M_Z is, from (52),

$$\begin{aligned} (E\{Z\})_i &\equiv \left(\int_{R^n} \frac{\nabla_x p(Y)}{p(Y)}\, p(Y)\, dY\right)_i = \int_{R^n} \frac{\partial}{\partial y_i}\, p(Y)\, dY \\ &= \int_{R^{n-1}} [p(Y)|_{y_i=-\infty}^{\infty}]\, dY = 0. \end{aligned} \tag{55}$$

The last equality holds when the density function satisfies (15). Thus the second term of (54) becomes, upon using (55) and integrating by parts,

$$\begin{aligned} &E\{(X - M_X)^T(Z - M_Z)\} \\ &= E\{X^TZ\} = \sum_{i=1}^{n} \int_{R^n} y_i \frac{\partial}{\partial y_i}\, p(Y)\, dY \\ &= \sum_{i=1}^{n} \left\{\int_{R^{n-1}} [y_i p(Y)|_{y_i=-\infty}^{+\infty}]\, dY - \int_{R^n} p(Y)\, dY\right\} \\ &= -n \end{aligned} \tag{56}$$

where we have assumed

$$\lim_{\|X\| \to \infty} Xp(X) = 0. \tag{57}$$

Using (55) and (56) in (54), we obtain

$$\begin{aligned} &E\{(Y - M_Y)^T(Y - M_Y)\} \\ &= E\{(X - M_X)^T(X - M_X)\} - 2an + a^2E\{Z^TZ\}. \end{aligned} \tag{58}$$

Thus for convergence in a mean-square sense all that is required is that a be chosen small enough; in particular, from (58),

$$0 < a < \frac{2n}{E\{Z^TZ\}}. \tag{59}$$

The optimum a, optimum in the sense of decreasing (54) as much as possible, is obtained by differentiating (58) with respect to a and noting that $E\{Z^TZ\}$ is greater than zero to obtain

$$a_{\text{opt}} = \frac{n}{E\{Z^TZ\}}. \tag{60}$$

This analysis gives us some insight into the choice of the parameter a. If a is too small, then the observations move only slightly towards the modes, but the variance decreases. If a is increased to the optimum value, then the step size is just right, and the variance decreases as much as possible. As a is made larger still, we begin to slightly overshoot the mode, but the variance still has decreased since it only involves distances from the mode. Finally, if a is made too large, we greatly overshoot the mode with the resulting variance being larger than that before the iteration. Thus one should be conservative in the choice of a so that although more iterations may be required to achieve tight clusters, at least the algorithm will not diverge.

4) The fourth reason for using the normalized gradient is that it can be estimated directly using the mean-shift estimate (41). When a Gaussian kernel function is used, the normalized gradient cannot be estimated directly, and we have to estimate both $\nabla_x p(X)$ and $p(X)$ separately and take their ratio.

B. Clustering Applications

As an example of gradient estimation in clustering, the algorithm of (45) was implemented on a computer using both the Gaussian kernel function approach and the sample mean-shift estimate (41) for $\nabla_x \ln p(X)$. The data for these experiments were computer-generated bivariate-normal random vectors. Sixty observations were obtained from each of three classes, each having identity covariance matrix and the following different means:

$$M_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad M_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix} \qquad M_3 = \begin{bmatrix} 0 \\ 4 \end{bmatrix}. \tag{61}$$

These distributions are slightly overlapping, the original data is shown in Fig. 2. By recursively applying the clustering algorithm (45) using gradient estimates, the data are transformed into three clusters. Fig. 3 shows the resulting transformed data for the Gaussian kernel estimate algorithm with $a = 0.5$ and $h = 0.8$, and for the mean-shift estimate algorithm with $a = 0.75$ and $h = 1.5$. Fig. 4 shows the resulting class assignments using the different algorithms. These results show that gradient estimation is indeed applicable to clustering problems.

The choice of the parameter h seems to depend upon the size of the clusters for which one is searching. This is due to the fact that the density estimate can be interpreted as approximating the convolution integral of $p(X)$ and $h^{-n}k(h^{-1}X)$, and thus the kernel function is acting as a smoothing filter on the density, see Fig. 5. Thus h determines the amount of smoothing of the density and correspondingly the elimination of modes that are too narrow or too close

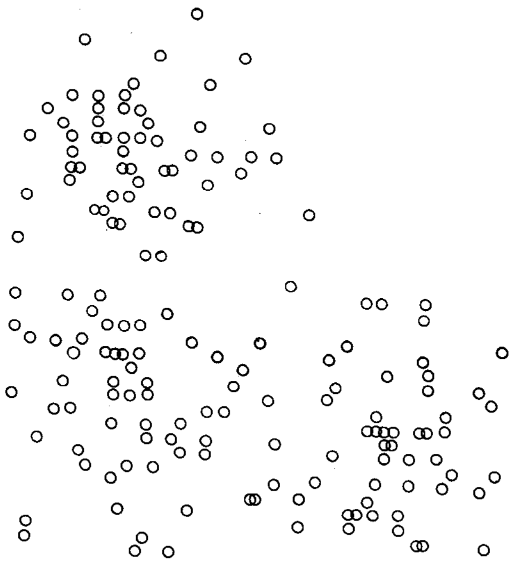

Fig. 2. Original data set.

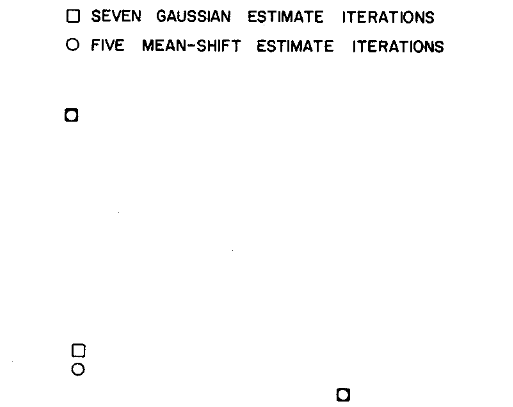

Fig. 3 Data set after clustering iterations.

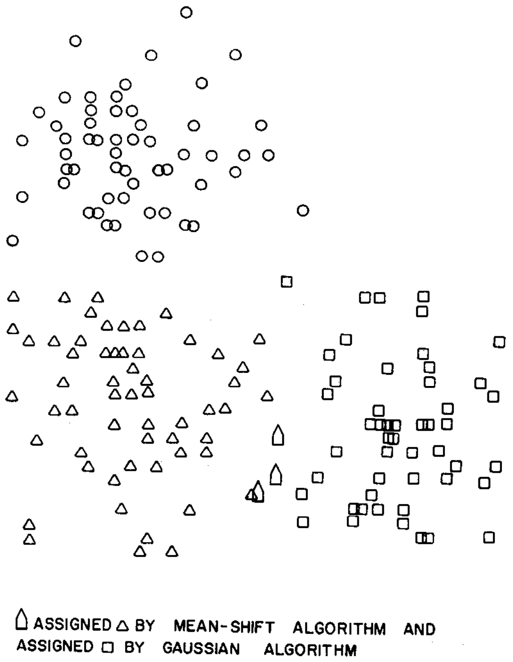

Fig. 4. Cluster assignments.

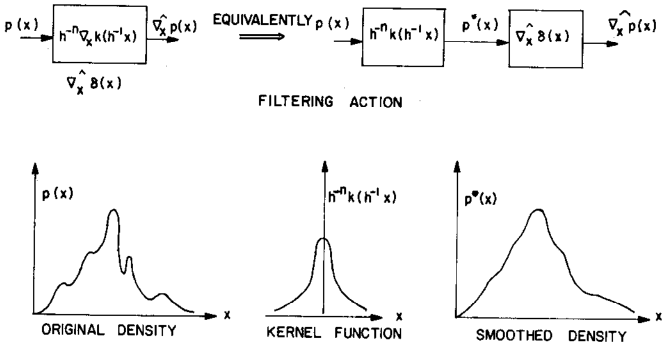

Fig. 5 Kernel smoothing action.

to other modes. See [4], [10], and [11] for discussions of the problem of the choice of h.

The choice of a that seemed to work well in these examples was

$$a = \frac{h^2}{n+2}. \tag{62}$$

Substituting (62) and (41) into (45) gives the mean-shift clustering algorithm

$$X_j^{i+1} = \frac{1}{k} \sum_{X_i^i \in S_h(X_j^i)} X_i^i. \tag{63}$$

This algorithm transforms each observation to the sample mean of the observations within the region S_h around it. Thus, if the entire data set is divided into convex subsets that are greater than distance h apart, the observations will always remain within their respective sets or clusters and cannot diverge. This is due to the fact that (63) is always a convex combination of members from the same convex set, therefore, it must also lie inside the set. Also, as soon as all the observations in such a set lie within a distance h of one another, the next iteration will transform them all to a common point, their sample mean.

C. Applications to Data Filtering

This approach can also be used as a data filter to reduce the effect of noise in determining the intrinsic dimensionality of a data set. The intrinsic dimensionality of a data set is defined (see [12]), to be the minimum number n_0 of parameters required to account for the observed properties of the data. The geometric interpretation is that the entire data set lies on a topological hypersurface of n_0 dimensions.

Fig. 6 shows a two-dimensional distribution that has intrinsic dimensionality of one. The Karhunen-Loève dominant eigenvector analysis (see [12]) used to find the number of principal axes of data variance of this distribution results in the two dominant axes shown in the figure. This suggests a dimensionality higher than the intrinsic value. This effect is avoided by using small local regions as in Fig. 7. Then the Karhunen-Loève analysis of these subsets indicates dimensionalities close to the intrinsic value.

This process experiences difficulty if there is noise present in the data. Thus if our observation space has $n > n_0$

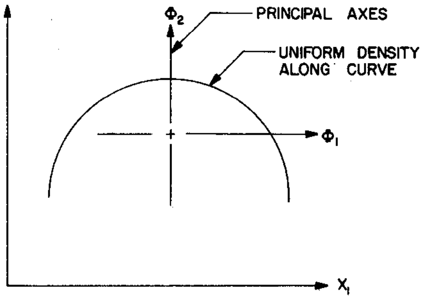

Fig. 6. Intrinsic dimensionality.

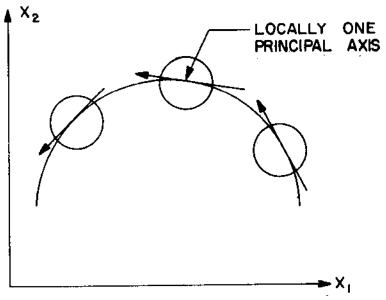

Fig. 7. Local data regions.

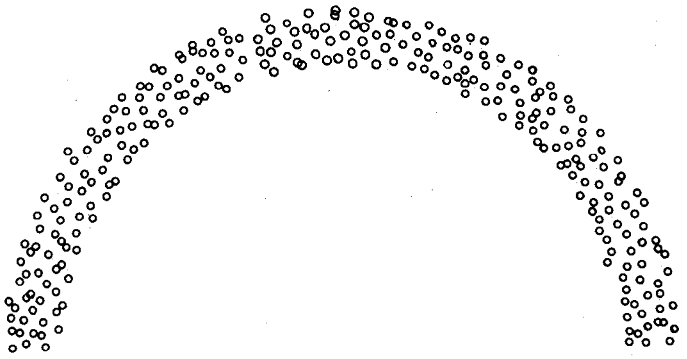

Fig. 8. Noisy data set.

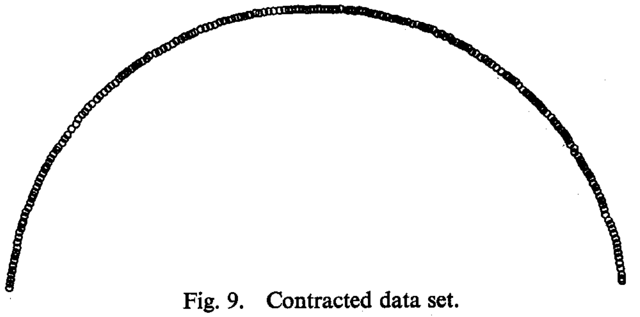

Fig. 9. Contracted data set.

dimensions, all of which contain noise in which the variance is of the order of the size of the local regions, then this process yields the value n as the intrinsic dimensionality. This effect could be offset by taking larger regions, but these regions then might include several surface convolutions, again resulting in an overestimate of the intrinsic dimensionality.

To reduce this effect, one would like to eliminate the noise in the minor dimensions, leaving only the dominant data surface. This can be achieved by using a few iterations of the mode clustering algorithm (45) as a data filter to contract the data onto its intrinsic data surface while still retaining the dominant properties of the data structure.

As an example, 400 observations were generated in the form

$$(x_1,x_2) = (r \cos \theta, r \sin \theta) \tag{64}$$

where r and θ were uniformly distributed over the regions

$$2.25 \leq r \leq 2.75 \qquad 0 \leq \theta \leq \pi \tag{65}$$

respectively. This noisy data set can be considered to have intrinsic dimensionality one, with the intrinsic data surface being a circle of radius r equal to 2.5. Fig. 8 shows the original noisy data set. By using two iterations of the transformation algorithm (45) with $h = 0.6$ and $a = h^2/3 = 0.12$, we see that the noise has been effectively removed leaving just the intrinsic data surface. This transformed data set is shown in Fig. 9. The results show that gradient estimation can be used to effectively eliminate the noise from a data surface.

V. Summary

By building upon previous results in nonparametric density estimation, we have been able to obtain a general class of kernel density gradient estimates. Conditions on the kernel function were derived to guarantee asymptotic unbiasedness, consistency, and uniform consistency of the estimate. By generalizing the results for a Gaussian kernel function, we developed a sample mean-shift estimate of the normalized gradient and extended it to a k-nearest-neighbor approach.

Applications of these estimates to the pattern recognition problems of clustering and intrinsic dimensionality determination were presented with examples. Like most nonparametric sample-based techniques, the direct application of the algorithm may be costly in terms of computer time and storage, but still the underlying philosophy of density gradient estimation in pattern recognition systems is worth investigating as a means of furthering our understanding of the pattern recognition process.

Appendix

A. Asymptotic Unbiasedness

In this section we will show that the gradient estimate is asymptotically unbiased for the density function that satisfies (15); that is,

$$\lim_{N\to\infty} E\{\hat{\nabla}_x p_N(X)\} = \nabla_x p(X). \tag{66}$$

Taking the expectation of $\hat{\nabla}_x p(X)$ of (9),

$$\begin{aligned} E\{\hat{\nabla}_x p_N(X)\} &= E\{h^{-n}\nabla_x k(h^{-1}(X - X_i))\} \\ &= h^{-n} \int_{R^n} \nabla_x k(h^{-1}(X - Y))p(Y)\, dY. \end{aligned} \tag{67}$$

Now looking at the ith component of this expectation and integrating by parts,

$$E\{(\hat{\nabla}_x p_N(X))_i\} = -h^{-n} \int_{R^{n-1}} k(h^{-1}(X - Y))p(Y)\Big|_{y_i=-\infty}^{+\infty} dY + h^{-n} \int_{R^n} k(h^{-1}(X - Y)) \frac{\partial}{\partial y_i} p(Y)\, dY. \tag{68}$$

Since $k(Y)$ is bounded, from (2), and $p(X)$ is a probability density function with (15) satisfied, the first term of (68) is zero as

$$\lim_{|y_i|\to\infty} |k(h^{-1}(X - Y))p(Y)| \le \sup_{Y\in R^n} |k(Y)| \lim_{|y_i|\to\infty} p(Y) = 0. \tag{69}$$

Under the conditions (2)–(6), we can apply (14) to obtain the second term

$$\lim_{N\to\infty} h^{-n} \int_{R^n} k(h^{-1}(X - Y)) \frac{\partial}{\partial y_i} p(Y)\, dY = \frac{\partial}{\partial x_i} p(X) \int_{R^n} k(Y)\, dY = \frac{\partial}{\partial x_i} p(X) \tag{70}$$

where the last equality follows from condition (5) on $k(X)$.

Asymptotic unbiasedness (16) now follows by substituting (69) and (70) in the limit of (68).

B. Consistency

In this section we will show that the gradient estimate is consistent in a mean-square sense; that is,

$$E\{\|\hat{\nabla}_x p_N(X) - \nabla_x p(X)\|^2\} = E\{\|\hat{\nabla}_x p_N(X) - E\{\hat{\nabla}_x p_N(X)\}\|^2\} + \|E\{\hat{\nabla}_x p_N(X)\} - \nabla_x p(X)\|^2 \tag{71}$$

goes to zero as N approaches infinity. Since the estimate is asymptotically unbiased for a density function that satisfies (15),

$$\lim_{N\to\infty} \|E\{\hat{\nabla}_x p_N(X)\} - \hat{\nabla}_x p(X)\|^2 = 0. \tag{72}$$

Substituting (9), the first term of (71) becomes

$$E\{\|\hat{\nabla}_x p_N(X) - E\{\hat{\nabla}_x p_N(X)\}\|^2\} = N^{-1} \sum_{i=1}^{n} \left[E\left\{\left[h^{-n} \frac{\partial}{\partial x_i} k(h^{-1}(X - Y))\right]^2\right\} - E^2\left\{h^{-n} \frac{\partial}{\partial x_i} k(h^{-1}(X - Y))\right\}\right]. \tag{73}$$

From (16), we have

$$\lim_{N\to\infty} E\left\{h^{-n} \frac{\partial}{\partial x_i} k(h^{-1}(X - Y))\right\} = \frac{\partial}{\partial x_i} p(X). \tag{74}$$

Now

$$E\left\{\left[h^{-n} \frac{\partial}{\partial x_i} k(h^{-1}(X - Y))\right]^2\right\} = h^{-2n} \int_{R^n} \left[\frac{\partial}{\partial z_i} k(h^{-1}Z)\right]^2 p(X - Z)\, dZ \tag{75}$$

$$= h^{-(n+2)} \left\{h^{-n} \int_{R^n} \left[\frac{\partial}{\partial(h^{-1}z_i)} k(h^{-1}Z)\right]^2 p(X - Z)\, dZ\right\}. \tag{76}$$

Using conditions (20)–(22), we see that $[k_i'(Y)]^2$ satisfies the conditions needed for (14) to hold. Thus

$$\lim_{N\to\infty} h^{-n} \int_{R^n} \left[\frac{\partial}{\partial(h^{-1}z_i)} k(h^{-1}Z)\right]^2 p(X - Z)\, dZ = p(X) \int_{R^n} [k_i'(Y)]^2\, dY \tag{77}$$

which is finite. Substituting (77), (76), and (74) into (73) and using condition (19) on $h(N)$, we obtain

$$\lim_{N\to\infty} E\{\|\hat{\nabla}_x p_N(X) - \nabla_x p(X)\|^2\} = \sum_{i=1}^{n} \left[\lim_{N\to\infty} (Nh^{n+2})^{-1} p(X) \int_{R^n} [k_i'(Y)]^2\, dY - \lim_{N\to\infty} N^{-1} \frac{\partial}{\partial x_i} p(X)\right] = 0. \tag{78}$$

C. Uniform Consistency

In this section we will show that the gradient estimate is uniformly consistent in probability; that is,

$$\lim_{N\to\infty} \Pr\{\sup_{X\in R^n} \|\hat{\nabla}_x p_N(X) - \nabla_x p(X)\| > \varepsilon\} = 0. \tag{79}$$

To show this, we notice that the gradient estimate is the convolution of the sample function and the function $h^{-n}\nabla_x k(h^{-1}X)$. Applying the properties of Fourier transforms, we see that the Fourier transform of $\hat{\nabla}_x p_N(X)$ is the product of the Fourier transforms of the sample function and that of $h^{-n}\nabla_x k(h^{-1}X)$. Using the inversion relationship, we see that

$$\hat{\nabla}_x p_N(X) = (2\pi)^{-1} \int_{R^n} \exp(-jW^T X)\Phi_N(W)[-jWK(hW)]\, dW \tag{80}$$

where W is an n-dimensional vector,

$$\Phi_N(W) \equiv \frac{1}{N} \sum_{i=1}^{N} \exp(jW^T X_i) \tag{81}$$

is the sample characteristic function,

$$K(W) \equiv \int_{R^n} \exp(jW^T X)k(X)\, dX \tag{82}$$

is the characteristic function of $k(X)$, and $-jWK(hW)$ is the characteristic function of $h^{-n}\nabla_x k(h^{-1}X)$ as in (28). Therefore,

$$E\{\sup_{R^n} \|\hat{\nabla}_x p_N(X) - E\{\hat{\nabla}_x p_N(X)\}\|\} \le (2\pi)^{-1} \int_{R^n} \|WK(hW)\| E\{|\Phi_N(W) - E\{\Phi_N(W)\}|\}\, dW \tag{83}$$

$$\le (2\pi)^{-1} \int_{R^n} \|WK(hW)\| \operatorname{var}\{\Phi_N(W)\}^{1/2}\, dW \tag{84}$$

$$= (2\pi)^{-1} \int_{R^n} \|WK(hW)\| [N^{-1} \operatorname{var}\{\exp(jW^T X)\}]^{1/2}\, dW \tag{85}$$

$$\le (2\pi)^{-1} \int_{R^n} \|WK(hW)\| N^{-1/2}\, dW \tag{86}$$

$$= (2\pi N^{1/2} h^{N+1})^{-1} \int_{R^n} \|-jWK(W)\|\, dW. \tag{87}$$

Since $-jWK(W)$ is absolutely integrable

$$\lim_{N\to\infty} E\{\sup_{R^n} \|\hat{\nabla}_x p_N(X) - E\{\hat{\nabla}_x p_N(X)\}\|\} = 0. \tag{88}$$

We can modify the asymptotically unbiased argument, taking into consideration the uniform continuity of $\nabla_x p(X)$, to obtain easily

$$\lim_{N\to\infty} [\sup_{R^n} \|E\{\hat{\nabla}_x p_N(X)\} - \nabla_x p(X)\|] = 0. \tag{89}$$

Now, applying the triangle inequality, we have

$$\sup_{R^n} \|\nabla_x p_N(X) - \nabla_x p(X)\| \leq \sup_{R^n} \|\hat{\nabla}_x p_N(X) - E\{\hat{\nabla}_x p_N(X)\}\| + \sup_{R^n} \|E\{\hat{\nabla}_x p_N(X)\} - \nabla_x p(X)\|. \tag{90}$$

Therefore, using (88) and (89) in the limit of (90), we obtain

$$\lim_{N\to\infty} E\{\sup_{R^n} \|\hat{\nabla}_x p_N(X) - \nabla_x p(X)\|\} = 0 \tag{91}$$

which implies by Markov's inequality

$$\lim_{N\to\infty} \Pr\{\sup_{R^n} \|\nabla_x p_N(X) - \nabla_x p(X)\| > \varepsilon\} = 0, \quad \text{for all } \varepsilon > 0. \tag{92}$$

References

[1] E. Fix and J. L. Hodges, "Discriminatory analysis, nonparametric discrimination," USAF Sch. Aviation Medicine, Randolph Field, Tex. Proj. 21-49-004, Rep. 4, Contr. AF-41-(128)-31, Feb. 1951.

[2] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Statist.*, vol. 27, pp. 832–837, 1956.

[3] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, pp. 1065–1067, 1962.

[4] T. Cacoullos, "Estimation of a multivariate density," *Ann. Inst. Statist. Math.*, vol. 18, pp. 179–189, 1966.

[5] E. A. Nadaraya, "On nonparametric estimates of density functions and regression curves," *Theory Prob. Appl.* (USSR), vol. 10, pp. 186–190, 1965.

[6] J. Van Ryzin, "On strong consistency of density estimates," *Ann. Math. Statist*, vol. 40, pp. 1765–1772, 1969.

[7] P. K. Bhattacharyya, "Estimation of a probability density function and its derivatives," *Sankhya: Ind. J. Stat.*, vol. 29, pp. 373–382, 1967.

[8] E. F. Schuster, "Estimation of a probability density function and its derivatives," *Ann. Math. Statist.*, vol. 40, pp. 1187–1195, 1969.

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

[10] V. A. Epanechnikov, "Nonparametric estimation of a multivariate probability density," *Theory Prob. Appl.* (USSR), vol. 14, pp. 153–158, 1969.

[11] M. Woodroofe, "On choosing a delta-sequence," *Ann. Math. Statist.*, vol. 41, pp. 1665–1671, 1970.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic, 1972, ch. 10.

A Finite-Memory Deterministic Algorithm for the Symmetric Hypothesis Testing Problem

B. CHANDRASEKARAN, MEMBER, IEEE, AND CHUN CHOON LAM

Abstract—**A class of irreducible deterministic finite-memory algorithms for the symmetric hypothesis testing problem is studied. It is shown how members of this class can be constructed to give a steady-state probability of error that decreases asymptotically faster in the number of states than the best previously known deterministic algorithm.**

Introduction

Let $X_1,X_2,\cdots$ be a sequence of independent identically distributed Bernoulli random variables with possible values H and T such that $\Pr(X_i = H) = p$. Consider the following symmetric hypotheses:

$$H_0: p = p_0$$

$$H_1: p = q_0 \equiv 1 - p_0$$

where, without loss of generality, $\frac{1}{2} < p_0 < 1$. We further assume that the hypotheses have equal prior probabilities. Let the data be summarized by a $2m$-valued statistic T that is updated according to the rule

$$T_n = f(T_{n-1},X_n), \qquad T_n \in \{1,2,\cdots,2m\}$$

where T_n is the value of T at time n. Let the decision d_n taken at time n be

$$d_n = d(T_n), \qquad d_n \in \{H_0,H_1\}.$$

For given f and d the probability of error is defined to be

$$P = E\left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} e_i\right)$$

where $e_i = 0$ or 1 depending on whether d_i is the correct decision or not. Hellman and Cover [1] have shown that a lower bound for P_e is

$$P_e{}^* = \left[1 + \left(\frac{p_0}{q_0}\right)^{2m-1}\right]^{-1}, \tag{1}$$

Manuscript received September 28, 1973; revised June 27, 1974. This work was supported by AFOSR Grant 72-2351.

B. Chandrasekaran is with the Department of Computer and Information Science, Ohio State University, Columbus, Ohio 43210.

C. Lam is with the Division of Statistics, Ohio State University, Columbus, Ohio 43210.