dence has been to examine some sources of confusion that arise in the practical uses of such relations and to show how the logical subtleties giving rise to the confusion may be dealt with properly. These subtleties occur over and over again in real-world model building and a selfconscious awareness of them is essential to the practice of good science in practical applications.

References

[1] J. N. Warfield, *Societal Systems*. New York: Wiley, 1976.

A Convexity Testing Method for Cluster Analysis

CHRISTIAN P. A. VASSEUR AND JACK G. POSTAIRE

Abstract—**The theory of convexity is used to develop a new mode detection procedure in an unsupervised context. The modes are detected by testing the convexity of the unknown underlying probability density function. The procedure is based on the analysis of the variations of the mean probability density defined over expanding observation domains of the pattern space. The estimates of the modes are then used to obtain separating surfaces which partition the data set into unimodal subsets. The resulting algorithm is governed by only one parameter, requires no initial classification, and also determines the number of clusters. Its performances are tested with artificially generated sets of data. The results corroborate the ability of the algorithm to identify clusters of various shapes and sizes in a very short computation time.**

I. Introduction

The last ten years have witnessed extensive activity in unsupervised learning procedures for the design of automatic classifiers [1], [2]. Many cluster detection algorithms have been developed based on the fundamental assumption that the patterns are drawn from a multimodal probability density function, each mode corresponding to one cluster. In this kind of approach, the discovery of the modes of the underlying distribution can be considered as the key problem.

Intuitively, the notion of convexity seems well-suited for cluster analysis because, as is shown later, the modes can be characterized by the convexity of the underlying probability density function. Aware of the fact that each new method offers additional useful information about the structure and the peculiarities of the data, the authors have been led to apply the properties of convex functions to cluster analysis.

A new mode seeking procedure is proposed under the assumption that the only information concerning the distribution is that which can be inferred from the input patterns. The procedure makes use of the notion of mean probability density (MPD) determined within expanding observation domains. The MPD is considered as a function of the size of the observation domain within which it is computed. The analysis of the variations of this function for growing domains centered at the samples gives pertinent information about the local convexity of the underlying probability density function (pdf). An aggregation procedure is then developed to cluster the samples where the underlying pdf is shown to be convex and which belong to the same class.

A formal statement of the method as well as proofs of relevant properties of the mean probability density for a specific class of domains are presented in Section II. These properties are used in Section III to develop the clustering algorithm. The mode seeking procedure finds subsets of samples which can be considered as the nuclei of the different classes. However, the procedure does not actually classify all the data. All samples left over can be assigned to their appropriate class by means of a classical supervised technique.

Finally the method is applied in Section IV to three artificially generated sets of data. Computation time is also given in relation to the dimensionality of the data.

II. Mode Separation

By measuring a number F of features, a pattern of a given set Γ of N elements can be represented by a point P of an Euclidian space E^F. It is assumed that one has no information about the set Γ other than these N multidimensional observations.

Many methods have been proposed to estimate a continuous probability density function from the input patterns. The parametric approach is based on the assumption that the form of the underlying density function is known [3].

When the patterns are not assumed to follow any parametric density, nonparametric procedures can be used to estimate the unknown distribution [4], [5].

Considering the shape of the probability density function estimated either by parametric or by nonparametric methods, clusters can be characterized by the convexity of the underlying density function. Clusters are no longer considered as local concentrations of patterns. They are now considered in terms of convexity.

Based on this property of convexity a mode seeking procedure is developed, using observation domains as defined below.

Definition of the Observation Domains (Fig. 1)

Let D_r be a convex reference domain in E^F, symmetrical with respect to its center and bounded by a hypersurface S_r. A family of domains denoted $D(P,\alpha)$ homothetic to the reference domain D_r centered at P is associated to each point P of E^F. The homothety is defined by its center P and a positive ratio α.

Each domain $D(P,\alpha)$ is referred to as an observation domain in the following and its boundary is denoted $S(P,\alpha)$.

Definition of the Mean Probability Density (MPD)

Let us consider a probability density function $f(X)$ defined at any point X of E^F. The function $f(X)$ can be considered as mass density. Thus, the probability for a point X to fall within the observation domain $D(P,\alpha)$ can be written as [6]:

$$\text{mass}\{D(P,\alpha)\} = \Pr\{X \in D(P,\alpha)\} = \int_{D(P,\alpha)} f(X)\,dX. \quad (1)$$

From the above, the mean probability density $\rho\{D(P,\alpha)\}$ can be defined as the ratio of the probability mass to the volume of the domain in which it was determined;

$$\rho\{D(P,\alpha)\} = \frac{\text{mass}\{D(P,\alpha)\}}{\text{vol}\{D(P,\alpha)\}}. \quad (2)$$

The fundamental property of the MPD which is used for mode detection is given by the following theorem, where $\rho\{D(P,\alpha)\}$ is considered as a function of the parameter α.

Fundamental Theorem

Let R be a region of the pattern space E^F in which the probability density function $f(X)$ is convex. Then, the MPD $\rho\{D(P,\alpha)\}$ determined within the observation domain $D(P,\alpha)$ is an upper bounded decreasing function of α, for nonnegative α such that $D(P,\alpha) \subset R$.

Manuscript received February 2, 1978; revised December 18, 1978, July 17, 1979, and November 29, 1979.

The authors are with the Centre d'Automatique, Bâtiment P2, Université de Lille 1, B.P. 36, 59650 Villeneuve D'Ascq, France.

0018-9472/80/0300-0145$00.75 ©1980 IEEE

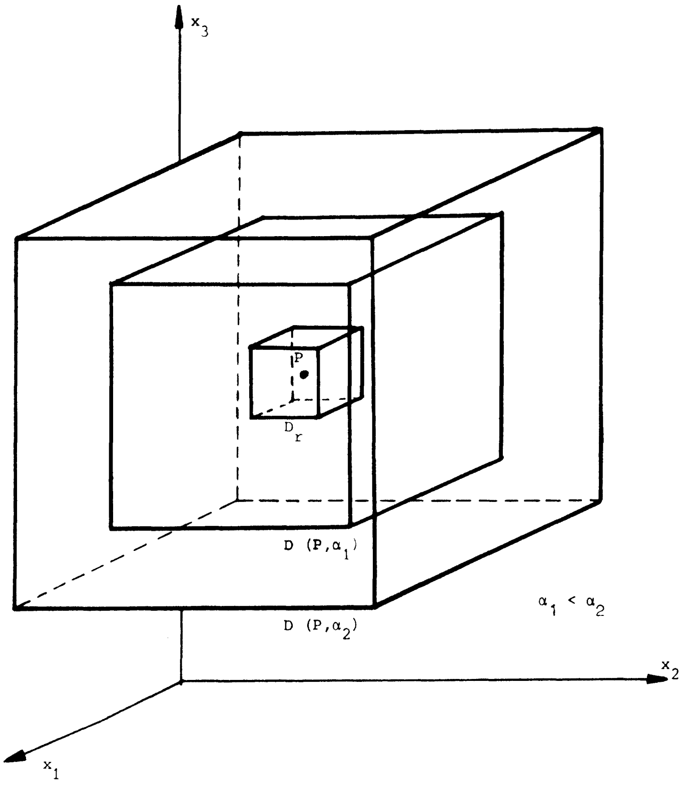

Fig. 1. Family of hypercubic observation domains in a 3-dimensional pattern space.

Similarly, if the probability function is concave in the region R, then the MPD $\rho\{D(P,\alpha)\}$ is, under the same conditions, a lower bounded increasing function of α.

A proof of this theorem is given in the Appendix.

Basis of the Mode Seeking Procedure

To each point P of the pattern space is associated an expanding observation domain $D(P,\alpha)$ over which the MPD is computed for positive ascending values of α, α starting from zero.

Let us assume that the point P does not stand on a boundary between a convex region and a concave region of the probability density function $f(X)$. Thus, there exists a neighborhood D_P of point P where the convexity of $f(X)$ is constant. In these conditions, there exists an $\alpha_0 \geqslant 0$ such that

$$D(P,\alpha) \subset D_P, \qquad \text{if } 0 \leqslant \alpha \leqslant \alpha_0.$$

Considering the fundamental theorem, we can conclude that if the MPD determiner within the observation domains $D(P,\alpha)$ is a descending function of α for $0 \leqslant \alpha \leqslant \alpha_0$ then $f(X)$ is convex in the neighborhood $D(P,\alpha_0)$ of P. Otherwise it is concave.

Hence the study of the variations of the MPD associated with an expanding observation domain provides the basis for a test of convexity. It is now theoretically possible to know the convexity of an unknown probability density function in the neighborhood of any point of the pattern space and consequently to detect the modes.

Clusters are then obtained by groupings of convex neighborhoods that are not separated by concave neighborhoods. By convex or concave neighborhoods, we mean neighborhoods wherein the pdf is estimated convex or concave.

III. The Computer-Oriented Clustering Algorithm

Let the given data set Γ consist of N data points X_i, each denoted by the vector:

$$X_i = (x_{i1}, x_{i2}, \text{——}, x_{iF})^T$$

where $i = 1, 2, \cdots, N$, and T indicates the transpose.

The data set is first normalized such that the data points lie within an hypercubic region of side length one.

It must be noted that the cluster separability will be influenced by this preliminary transformation. However, the study of more sophisticated transformations and the subsequent improvement of class separability falls outside the scope of this correspondence.

The selected observation domains are hypercubes centered at the data points. The choice of hypercubes is based on computation time considerations. As an example, if the reference domain D_r is an hypersphere, then the tests of convexity require the calculation of euclidean distances. On the other hand, if the reference domain D_r is an hypercube, the tests require only comparisons between coordinates and this procedure is faster and simpler than the preceeding technique. From a practical point of view, the direction of variations of the MPD for an observation domain increasing around X_i is determined by estimating the MPD within two hypercubic neighborhoods of X_i, denoted $D_1(X_i)$ and $D_2(X_i)$ centered at X_i, respectively of volume V_1 and V_2, with $V_2 > V_1$.

Let k_1 and k_2 denote the number of samples falling respectively in $D_1(X_i)$ and $D_2(X_i)$. The estimated $\hat{\rho}\{D_1(X_i)\}$ and $\hat{\rho}\{D_2(X_i)\}$ of the MPD within $D_1(X_i)$ and $D_2(X_i)$ are given by

$$\hat{\rho}\{D_1(X_i)\} = \frac{k_1}{V_1} \quad \text{and} \quad \hat{\rho}\{D_2(X_i)\} = \frac{k_2}{V_2}.$$

If $\hat{\rho}\{D_1(X_i)\}$ is greater than $\hat{\rho}\{D_2(X_i)\}$, the pdf is estimated convex in the neighborhood $D_1(X_i)$ of X_i. Otherwise, the pdf is estimated concave in this neighborhood.

We have chosen a k-nearest neighbor procedure as described in [7] to estimate the MPD in $D_1(X_i)$. The hypercubic domain $D_1(X_i)$ is grown until it encloses k_1 neighbors of X_i, where $k_1 = k_0\sqrt{N}$. Then the homothetic domain $D_2(X_i)$ is adjusted such as $V_2 = \lambda.\ V_1$ (choosing $\lambda = 2^F$ has led to good results).

Once all the tests of convexity have been achieved, the clusters are generated by grouping the overlapping neighborhoods in which the pdf has been shown to be convex. Two neighborhoods are assigned to the same cluster if they have at least one common sample.

When this aggregation procedure is completed, all points belonging to convex regions can be considered as the nuclei of the classes for the classification of the remaining unassigned samples.

IV. Experimental Results

The performance of the algorithm depends to a large extent on the choice of the parameter k_0 which controls the number of samples to be captured by the observations domains $D_1(X_i)$. *A priori*, without any additional information, any choice of k_0 is as good as any other. However this parameter can be considered as a smoothing parameter. The greater k_0, the less the procedure is sensitive to small variations of convexity of the pdf. Choosing this kind of parameter in the middle of the largest range where the number of detected modes remains constant has been shown to be a good procedure to optimize a number of algorithms [8].

The performances of the present algorithm are tested on three artificially generated sets of data similar to those in [9].

Experiment 1: A hundred pseudorandom samples are taken from each of the following three bivariate Gaussian distributions (M_i denotes the mean and Σ_i the covariance matrix for distribution i) (Fig. 2(a));

$$M_1 = [0 \quad 0]^T \quad M_2 = [0 \quad 3]^T \quad M_3 = [3 \quad 0]^T$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

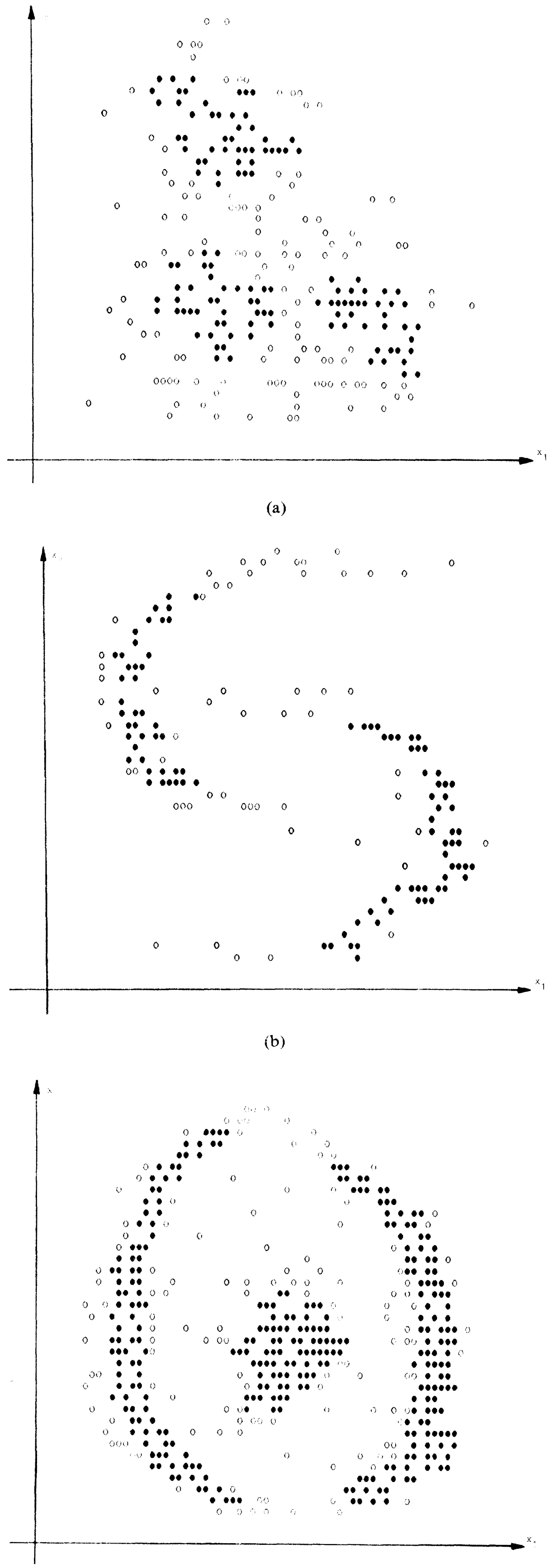

Fig. 2. Results of experiments 1, 2, and 3 with $k_0=0.6$; ○: samples constituting the data set, ●: samples constituting the nuclei of the detected modes.

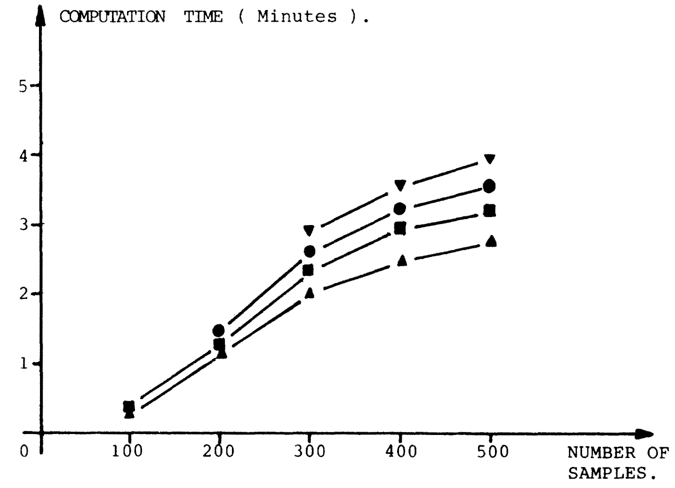

Fig. 3. Computation time; ▲ 2-dimensional data set, ■ 3-dimensional data set, ● 4-dimensional data set, ▼ 5-dimensional data set.

Experiment 2: The algorithm was applied to the set of nonspherical bivariate data of the Fig. 2(b). This set consists of 155 samples.

Experiment 3: The method was tested on a strongly nonlinearly separable set of 480 samples shown in Fig. 2(c).

Figs. 2(a), 2(b), and 2(c) show for each set of data, the nuclei of the classes detected by the algorithm. Note that, to speed up the procedure, the convexity test is not performed systematically around each sample. In fact, experiment show that testing the convexity only around samples which have not previously been shown to belong to convex neighborhoods leads to good results.

Fig. 3 indicates the algorithm's computation time as a function of both the sample size and the dimensionality of the data (all simulations were performed on a PDP 11-34 computer). For each experiment the samples are taken from three spherical normal equiprobable distributions with mean vectors M_1, M_2, M_3 and equal covariance matrices $\Sigma_1=\Sigma_2=\Sigma_3$ as shown in Table I.

The running time is seen to rise somewhat slower than a linear function of the sample size. This property is the result of the procedure which speeds up the algorithm by reducing the number of convexity tests, especially in regions of high density of samples. Moreover, Fig. 3 shows that for a given number of samples, the running time is nearly a linear function of the dimensionality.

Note that the sample size required to detect the modes increases with the dimensionality of the data. This is a well-known practical limitation on the k-nearest neighbor estimation scheme. Fortunately the convexity testing method does not require a very accurate estimation of the MPD's, so that the procedure remains practicable even for relatively small sample sizes. For the examples of Fig. 3, the mode detection procedure requires 60 and 90 samples per class for the 4- and the 5-dimension cases, respectively.

V. Conclusion

The mode detection procedure discussed in this correspondence is based on the determination of the local convexity of the underlying pdf from the samples. Modes are identified as convex regions of the pdf and can virtually be of any shape and size.

The algorithm, which is noniterative, requires neither a starting classification, nor an *a priori* number of clusters. Only one parameter needs to be adjusted.

For each detected mode. the algorithm determines a nucleus which can be used as a training set for learning the separating surfaces between the clusters.

The computation time required by the method is linearly dependent on the number of samples and the dimensionality of the data. The method therefore remains computationally efficient for the data of high dimensionality.

TABLE I

2-DIMENSIONAL DATA SET. $M_1 = \begin{bmatrix} 8 \\ 8 \end{bmatrix}, M_2 = \begin{bmatrix} 14 \\ 14 \end{bmatrix}, M_3 = \begin{bmatrix} 14 \\ 2 \end{bmatrix}$ $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$	3-DIMENSIONAL DATA SET. $M_1 = \begin{bmatrix} 8 \\ 8 \\ 8 \end{bmatrix}, M_2 = \begin{bmatrix} 14 \\ 14 \\ 14 \end{bmatrix}, M_3 = \begin{bmatrix} 14 \\ 14 \\ 2 \end{bmatrix}$ $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$
4-DIMENSIONAL DATA SET. $M_1 = \begin{bmatrix} 8 \\ 8 \\ 8 \\ 8 \end{bmatrix}, M_2 = \begin{bmatrix} 14 \\ 14 \\ 14 \\ 14 \end{bmatrix}, M_3 = \begin{bmatrix} 14 \\ 14 \\ 14 \\ 2 \end{bmatrix}$ $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$	5-DIMENSIONAL DATA SET. $M_1 = \begin{bmatrix} 8 \\ 8 \\ 8 \\ 8 \\ 8 \end{bmatrix}, M_2 = \begin{bmatrix} 14 \\ 14 \\ 14 \\ 14 \\ 14 \end{bmatrix}, M_3 = \begin{bmatrix} 14 \\ 14 \\ 14 \\ 14 \\ 2 \end{bmatrix}$ $\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix}$

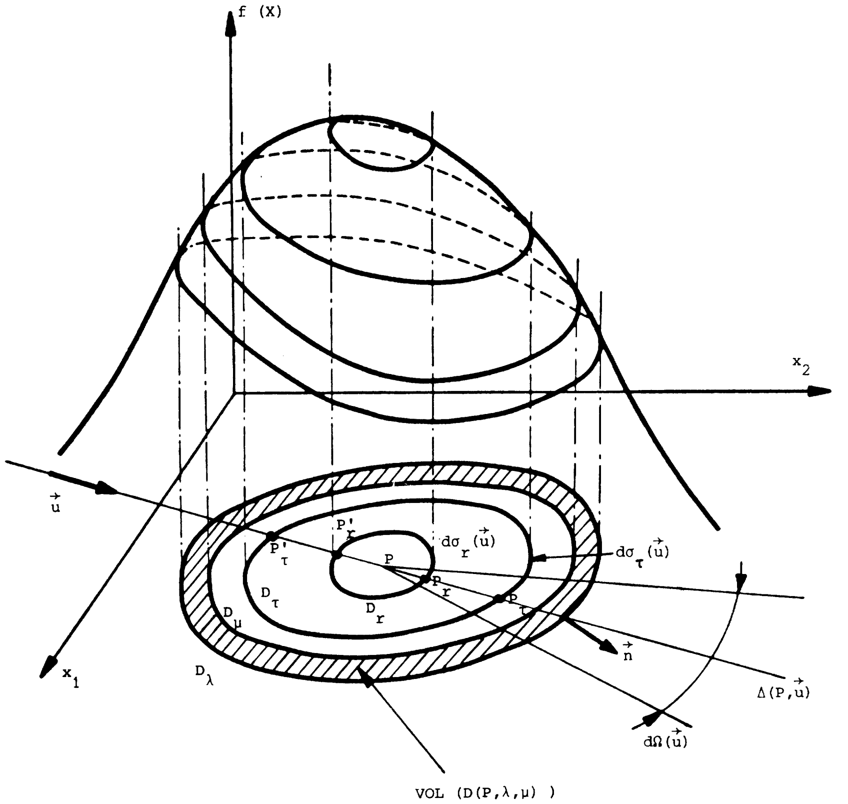

Fig. 4. A pdf and the associated observation domains for $F=2$.

Satisfactory results have been obtained from computer simulation, showing that the concept of convexity is well-suited to cluster analysis.

Appendix

Proof of the Fundamental Theorem for f(X) Convex (cf. Fig. 4)

A) Consider the variation $\delta\rho$ of the MPD due to an increment $\delta\alpha$ of the ratio of homothety α:

$$\delta\rho=\rho\{D(P,\alpha+\delta\alpha)\}-\rho\{D(P,\alpha)\}.$$

Introducing the domain $D(P,\alpha,\alpha+\delta\alpha)$ enclosed between $D(P,\alpha+\delta\alpha)$ and $D(P,\alpha)$ such that

$$\operatorname{mass}\{D(P,\alpha+\delta\alpha)\}=\operatorname{mass}\{D(P,\alpha)\}+\operatorname{mass}\{D(P,\alpha,\alpha+\delta\alpha)\},$$

and

$$\operatorname{volume}\{D(P,\alpha+\delta\alpha)\}=\operatorname{volume}\{D(P,\alpha)\}+\operatorname{volume}\{D,\alpha,\alpha+\delta\alpha\},$$

one gets

$$\delta\rho=\frac{\operatorname{volume}\{D(P,\alpha,\alpha+\delta\alpha)\}}{\operatorname{volume}\{D(P,\alpha+\delta\alpha)\}}(\rho\{D(P,\alpha,\alpha+\delta\alpha)\}-\rho\{D(P,\alpha)\}). \tag{A1}$$

It is clear that the sign of $\delta\rho$ is equal to the sign of $\delta'\rho$ such that

$$\delta'\rho=\rho\{D(P,\alpha,\alpha+\delta\alpha)\}-\rho\{D(P,\alpha)\}.$$

$\delta'\rho$ is the difference between the MPD within the "crust" of the observation domain and the MPD within the domain itself.

B) Let us now evaluate the sign of $\delta'\rho$. Let $\Delta(P,\vec{u})$ be a straight line defined by the unit vector $\vec{u}$ and by the point P which is both the center of symmetry and the center of homothety of the family of observations domains $D(P,\alpha)$. This line crosses the boundary S_r of D_r at two points P_r and P_r' and the boundary $S(P,\tau)$ of $D(P,\tau)$ at two points P_τ and P_τ' such that

$$\vec{PP}_r=-\vec{PP}_r'=1(\vec{u})\cdot\vec{u}$$

$$\vec{PP}_\tau=-\vec{PP}_\tau'=\tau\cdot 1(\vec{u})\cdot\vec{u}$$

where the positive scalar $1(\vec{u})$ is defined by the geometrical properties of D_r.

Two homothetical infinitesimal elements of the surfaces S_r and $S(P,\tau)$ respectively centered at P_r and P_τ and denoted $d\sigma_r$ and $d\sigma_\tau$ verify the following relation:

$$d\sigma_\tau=\tau^{F-1}\cdot d\sigma_r$$

where F is the dimensionality of the feature space.

If now $d\omega_\tau$ denotes the infinitesimal element of volume generated by an elementary displacement of $d\sigma_\tau$ induced by a variation $d\tau$ of τ such that

$$d\vec{PP}_\tau=1(\vec{u})\cdot\vec{u}\cdot d\tau,$$

one can write

$$d\omega_\tau=1(\vec{u})\cdot(\vec{u}x\vec{n})\cdot d\sigma_r\cdot\tau^{F-1}\cdot d\tau$$

where $\vec{n}$ denotes the outward pointing normal to S_r at P_r and $\vec{u}x\vec{n}$ designates the scalar product of $\vec{u}$ and $\vec{n}$.

The probability mass within the domain $D(P,\alpha)$ can be written as

$$\operatorname{mass}\{D(P,\alpha)\}=\int_0^\alpha\left\{\int_{S_r}f(P_\tau)\cdot 1(\vec{u})\cdot(\vec{u}x\vec{n})\cdot d\sigma_r\right\}\cdot\tau^{F-1}\cdot d\tau. \tag{A2}$$

Since the hypersurface S_r is symmetrical with respect to P, (A2) can be rewritten as

$$\operatorname{mass}\{D(P,\alpha)\}=\frac{1}{2}\int_0^\alpha\left\{\int_{S_r}(f(P_\tau)+f(P_\tau'))\cdot 1(\vec{u})\cdot(\vec{u}x\vec{n})\cdot d\sigma_r\right\}\cdot\tau^{F-1}\cdot d\tau. \tag{A3}$$

In any direction $\vec{u}$, the property of symmetry makes possible to write

$$P_\tau=p\cdot P_\alpha+q\cdot P_\alpha',$$
$$P_\tau'=q\cdot P_\alpha+p\cdot P_\alpha'$$

where p and q are two positive real numbers such that $p+q=1$.

Assuming that the probability density function $f(X)$ is convex in a domain D which contains $D(P,\alpha)$ one can write [10]

$$f(P_\tau)=f(p\cdot P_\alpha+q\cdot P_\alpha')\geqslant p\cdot f(P_\alpha)+q\cdot f(P_\alpha') \tag{A4}$$

and similarly,

$$f(P_\tau')=f(q\cdot P_\alpha+p\cdot P_\alpha')\geqslant q\cdot f(P_\alpha)+p\cdot f(P_\alpha') \tag{A5}.$$

Hence, by summing (A4) and (A5)

$$f(P_\tau)+f(P_\tau')\geqslant f(P_\alpha)+f(P_\alpha').$$

From (A3)

$$\operatorname{mass}\{D(P,\alpha)\}\geqslant\frac{K(\alpha)}{2}\cdot\int_0^\alpha\tau^{F-1}d\tau \tag{A6}$$

where

$$K(\alpha)=\int_{S_r}\{f(P_\alpha)+f(P_\alpha')\}\cdot 1(\vec{u})\cdot(\vec{u}x\vec{n})\cdot d\sigma_r.$$

From the properties of homothety we have

$$\operatorname{volume}\{D(P,\alpha)\}=\alpha^F\operatorname{volume}\{D_r\}$$

so that the relation (A6) yields

$$\frac{\operatorname{mass}\{D(P,\alpha)\}}{\operatorname{volume}\{D(P,\alpha)\}}\geqslant\frac{K(\alpha)}{2F\cdot\operatorname{volume}\cdot(D_r)}. \tag{A7}$$

Similarly, for the domain $D(P,\alpha,\alpha+\delta\alpha)$, it can be shown that

$$\frac{\operatorname{mass}\{D(P,\alpha,\alpha+\delta\alpha)\}}{\operatorname{volume}\{D(P,\alpha,\alpha+\delta\alpha)\}}\leqslant\frac{K(\alpha)}{2F\cdot\operatorname{volume}(D_r)}.$$

From (A1), (A7), and (A8) one can conclude that $\delta\rho\leqslant 0$.

C) The same procedure holds good for a concave probability density function. In this case, we obtain $\delta\rho\geqslant 0$.

Acknowledgment

The authors wish to thank the referee for his helpful comments and suggestions.

References

[1] L. Kanal, "Pattern in pattern recognition: 1968–1974," *IEEE Trans. on Inform. Theory*, vol. IT-20, no. 6, Nov. 1974.

[2] Proceedings of the 3rd and 4th International Joint Conferences on Pattern Recognition—1976 and 1978.

[3] P. W. Cooper, "Nonsupervised learning in statistical pattern recognition." *Methodology of Pattern Recognition* New York: Academic, 1969, pp. 97–109.

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

[5] T. Cacoullos, "Estimation of a multivariate density," *Ann. Inst. Statist. Math.*, no. 18, p. 179, 1966.

[6] A. Papoulis, *Probability, Random Variables and Stochastic Processes.* New York: McGraw-Hill, 1965.

[7] T. M. Cover and B. P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. on Inform. Theory*, vol. IT-13 pp. 21–27, Jan. 1967.

[8] D. J. Eagen and F. R. Fromm, "Cluster analysis based on dimensional information with applications to feature selection and classification," *IEEE Trans. on Syst., Man, and Cybern.*, vol. SMC-4, no. 3, pp. 284–294, 1974.

[9] W. L. Koontz, P. M. Narendra, and K. Fukunaga, "A graph theoretic approach to nonparametric cluster analysis," *IEEE Trans. on Comp.*, vol. C-25, pp. 936–944, 1976.

[10] H. G. Eggleston, "Convexity," *Cambridge Tracts in Math. and Math. Physics*, 1963.