Cluster Mapping with Experimental Computer Graphics

EDWARD A. PATRICK, MEMBER, IEEE, AND FREDERIC P. FISCHER II, STUDENT MEMBER, IEEE

Abstract—The unsupervised estimation problem has been conveniently formulated in terms of a mixture density. It has been shown that a criterion naturally arises whose maximum defines the Bayes minimum risk solution. This criterion is the expected value of the natural log of the mixture density. By making the assumptions that the component densities in the mixture are truncated Gaussian, the criterion has a greatly simplified form. This criterion can be used to resolve mixtures when the number of classes as well as the class covariances are unknown. In this paper a technique is presented where an assumed test covariance is supplied by an experimenter who uses a test function as a "portable magnifying glass" to examine data. Because the experimenter supplies the covariance and thus the test function, the technique is especially suited for interactive data analysis.

Index Terms—Clustering, computer display of mixed data, computer graphics in pattern recognition, interactive data analysis, interactive pattern recognition system, mixture density, pattern recognition, sorting data unsupervised estimation of densities.

Introduction

THE UNSUPERVISED estimation problem is conveniently formulated in terms of mixtures and mixing parameters; let $x_1, x_2, \cdots, x_n$ be l dimensional vector samples where x has a density function $h(\mathbf{x}|\boldsymbol{B})$:

$$h(x|B) = \sum_{i=1}^{M} f(\mathbf{x}|\boldsymbol{\alpha}_i)P(\boldsymbol{\alpha}_i) \tag{1}$$

$$\boldsymbol{B} = \{\boldsymbol{\alpha}_i, P(\boldsymbol{\alpha})_i\}_{i=1}^{M} \tag{2}$$

where $\boldsymbol{B}$ is in the parameter space and $h(\mathbf{x}|\boldsymbol{B})$ is called a mixture. It has been shown by Patrick and Costello [12], [13] that a criterion naturally arises whose maximum defines the Bayes minimum risk solution. This criterion denoted $\eta(\boldsymbol{B})$ is the expected value of the natural logarithm of the mixture density of the observation vectors:

$$\eta(\boldsymbol{B}) = \int \ln h(\mathbf{x}|B)h(\mathbf{x}|\boldsymbol{B}^*)dx \tag{3}$$

where $\boldsymbol{B}^*$ is the true parameter point. They have also shown that it is possible to approximate the Bayes solution by using $\eta(\boldsymbol{B})$ as a regression function for stochastic approximation or for finding a partition of the sample space. The latter proceeds as follows: for an asymptotic minimum risk solution it is sufficient to find the parameter vector $\boldsymbol{B}$ that maximizes

$$\eta(\boldsymbol{B}) = \int \ln\left[\sum_{i=1}^{M} f(\mathbf{x}|\boldsymbol{\alpha}_k)P(\boldsymbol{\alpha}_k)\right]h(\mathbf{x})dx.$$

The sample space is partitioned into M disjoint regions where the regions are defined

$$S^k \triangleq \{x: f(\mathbf{x}|\boldsymbol{\alpha}_k)P(\boldsymbol{\alpha}_k) > f(\mathbf{x}|\boldsymbol{\alpha}_j)P(\boldsymbol{\alpha}_j) \text{ all } j \neq k\} \quad k = 1, 2, \cdots, M. \tag{4}$$

It is assumed that over each partitioned set the class density is Gaussian having mean vector $\boldsymbol{\Upsilon}_k$, covariance matrix $\boldsymbol{\Phi}_k$, with the density truncated at the partition boundary. The true mixture $h(\mathbf{x})$ is assumed bounded. Under these assumptions $\eta(\boldsymbol{B})$ can be expanded and reduced to

$$\eta(\boldsymbol{B}) = \sum_{i=1}^{M} P(\boldsymbol{\alpha}_i) \ln\left[\frac{P(\boldsymbol{\alpha}_i)}{(2\pi)^{l/2}|\boldsymbol{\Phi}_i|^{1/2}}\right] - \frac{1}{2}. \tag{5}$$

Then maximizing $\eta(\boldsymbol{B})$ is equivalent to finding the partition (4) which maximizes (5). Given the partition (4), the maximum likelihood estimate of the mean vector $\boldsymbol{\Upsilon}_k$ is the sample mean of the samples in the kth region of the partition.

If the covariances $\{\boldsymbol{\Phi}_i\}$ are known and the classes are "well separated," there is no problem in visually determining the samples belonging to a class for the two dimensional case. For data vectors of dimensionality higher than two, it is clear that data vectors close to a particular data vector $\mathbf{x}_j$ belong to the same class as does x_j. In this paper we take such a "clustering" approach to the problem which can be motivated by the criteria $\eta(\boldsymbol{B})$ under the assumptions that the covariances are known $\{\boldsymbol{\Phi}_i\}$ and separability. The approach is to define a cluster as a set of observations which "likely" originated from the same mixture component; thus a *local neighborhood* philosophy is utilized. Essentially, each observation $\mathbf{x}_j$ is associated with (mapped to) the parameters of a Guassian mixture component dominating in a neighborhood about $\mathbf{x}_j$. Hence, observations drawn from the *same mixture component* are mapped approximately to the *same point* in the parameter space. Clusters are identified by finding the subsets of the observations mapped to the same "fuzzy" point.

Manuscript received June 9, 1969; revised July 1, 1969. This work was supported by the Rome Air Development Center under Contract F 30 602-68-C-0186. This paper was presented at the IEEE Computer Group Conference, Minneapolis, Minn., June 17–19, 1969.

The authors are with the School of Electrical Engineering, Purdue University, Lafayette, Ind.

Whereas the quality of unsupervised estimation can be measured by a criteria $\eta(\boldsymbol{B})$, the *clustering* approach taken in this paper does not utilize such a criteria. The advantage of not using such a criteria when clustering is simplification, as well as the fact that the technique will always work when samples are "well separated" and the covariances known. A disadvantage of the clustering approach is that it can fail to separate classes when the above assumptions are violated; then a criteria like $\eta(\boldsymbol{B})$ should be used.

There are many approaches which can be called clustering as is indicated in a literature review by Ball [1]. The least complex solution may be that of Sebestyen [2], where clusters are defined to be the set of all points within a distance T of a cluster center. Clusters also can be defined in terms of a thresholded similarity matrix [3]. Then, a cluster may be the set of all points which can be connected together with links of length less than a threshold T. The entire information of the data set may be used to define a cluster, as did Ball and Hall when they considered finding the mean distance from the samples within a group to their mean. A partial listing of papers on clustering and unsupervised estimation are [1]–[7], [8], [10], [12], [13]. Pearson [4] appears to be the earliest work, and Nagy [14] has presented a recent review of pattern recognition including the clustering problem.

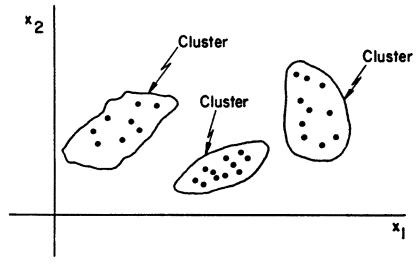

Fig. 1. Clustering.

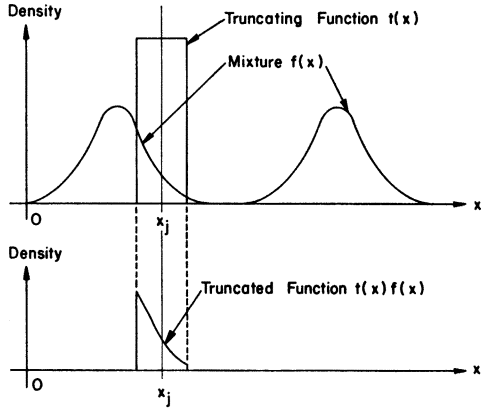

Fig. 2. 0–1 truncating function.

The Cluster Map

Although this approach will not be taken, an intuitive way to achieve the above results using a neighborhood philosophy is as follows: suppose that we are given a neighborhood about the point x_j. For example, the neighborhood may be the set of all points inside the circle drawn through the rth nearest neighbor to x_j (see Fig. 1). Let $t(\mathbf{x})$ equal one on the inside of this region, and let $t(\mathbf{x})$ equal 0 elsewhere. For this particular $t(\mathbf{x})$ and $\mathbf{x}_j$, assume that the dth mixture component dominates all other components inside this neighborhood. More specifically, assume that

$$\begin{aligned} h(x)t(x) &= t(x)\sum_{i=1}^{M} P_i f(\mathbf{x}\mid \boldsymbol{\Phi}_i, \boldsymbol{\gamma}_i) \\ &= t(\mathbf{x})P_d f(\mathbf{x}\mid \boldsymbol{\Phi}_d, \boldsymbol{\gamma}_d) \end{aligned} \tag{6}$$

where $f(\mathbf{x}\mid \boldsymbol{\Phi}_d, \boldsymbol{\Upsilon}_d)$ is a Gaussian mixture component dominant at $\mathbf{x}_j$. Thus, $h(\mathbf{x})t(\mathbf{x})$ is approximately a single, truncated, multivariate Gaussian function. The truncating function $t(\mathbf{x})$ removes the other mixture components from consideration, as is shown in Fig. 2. Since $h(\mathbf{x})t(\mathbf{x})$ is a truncated Gaussian density, a relationship between the moments of $f(\mathbf{x})t(\mathbf{x})$ and the moments of $h(x)$ might be determined. Using a sample density[1] $\hat{h}(x)$ instead of $h(\mathbf{x})$, estimates of the moments of $h(\mathbf{x})t(\mathbf{x})$ could be used to obtain estimates $\{\boldsymbol{\Phi}_{d_j}, \boldsymbol{\Upsilon}_{d_j}\}$. If the same procedure is carried out at each observation, observations under the same component would be associated with approximately the same estimates in the parameter space. It would then be a simple accomplishment to identify these tight "clusters" in the parameter space. Although conceptually promising, this approach is rejected for the reason that the relations between moments of a truncated multivariate Gaussian and the moments of the Gaussian are unknown, except for the univariate case [10]. To avoid this problem, another way is used to estimate the parameters of the mixture using moment estimators.

Instead of choosing a truncating function $t(x)$ which is one inside and zero outside a neighborhood, choose a better suited function—the ubiquitous multivariate Gaussian density function centered at $\mathbf{x}_j$ with covariance matrix[2] $\boldsymbol{\Sigma}$. We still may visualize a "pseudo-neighborhood" of $\mathbf{x}_j$ to be those points within the circle of concentration of the Gaussian function, $t(x) = t(\mathbf{x}\mid \boldsymbol{\Sigma}, x_j)$ with mean $\mathbf{x}_j$ and covariance $\boldsymbol{\Sigma}_j$. The choice of a Gaussian function leads to great mathematical simplicity.

Then using a theorem by Miller [11], on page 24,

$$t(\mathbf{x}\mid \boldsymbol{\Sigma}, \mathbf{x}_j)f(\mathbf{x}\mid \boldsymbol{\Phi}_j, \boldsymbol{\gamma}_d) = kf(\mathbf{x}\mid \boldsymbol{R}, \boldsymbol{C}) \tag{7a}$$

$$\boldsymbol{R} = (\boldsymbol{\Phi}_d^{-1} + \boldsymbol{\Sigma}^{-1})^{-1} \tag{7b}$$

$$\boldsymbol{C} = R(\Phi_d^{-1}\boldsymbol{\gamma}_d + \boldsymbol{\Sigma}^{-1}x_j)$$

$$k = \left(\frac{1}{2\pi}\right)^{n/2}\left(\mid R^{-1}\mid\mid \boldsymbol{\Phi}_d\mid\mid \boldsymbol{\Sigma}\mid\right)^{-1/2}$$

$$\cdot\exp\left\{\tfrac{1}{2}[\boldsymbol{C}^t\boldsymbol{R}^{-1}\boldsymbol{C} - \boldsymbol{\gamma}_d{}^t\boldsymbol{\Phi}_d^{-1}\boldsymbol{\gamma}_d - \mathbf{x}_j{}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}_j]\right\}.$$

[1] $\hat{h}(x) = 1/n\sum_{i=1}^{n}\delta(\mathbf{x}-\mathbf{x}_i)$; $\delta(\mathbf{x})$ is the Dirac delta.

[2] This covariance matrix $\boldsymbol{\Sigma}$ for the test function will be supplied interactively by the operator.

In other words, a "truncated" Gaussian mixture component is also Gaussian, and its mean and covariance matrix is given above in terms of the moments of the mixture component and the truncating function $t(\mathbf{x})$. (See Fig. 3.)

In particular, given the moments of the truncated function and the function $t(\mathbf{x})$, the inverse relationship is easily obtained:

$$\begin{aligned} \Phi_d &= (R^{-1} - \Sigma^{-1})^{-1} \\ \gamma_d &= (\Phi_d \Sigma^{-1} + I)(C - x_j) + x_j. \end{aligned} \tag{8}$$

Thus, by (6) and (7),

$$h(x)t(x) = P_d k f(x \mid R, C). \tag{9}$$

The preceding development is used in the following way: the moments of the function $h(x)t(x)$ are estimated using the moments of the function $\hat{h}(x)t(x)$ where $\hat{h}(x)$ is the sample density. That is, letting

$$\hat{m}_0 = \frac{1}{n} \sum_{s=1}^{n} t(x_s \mid \Sigma, x_j)$$

$$\hat{m}_\mu{}^1 = \frac{1}{m_0} \frac{1}{n} \sum_{s=1}^{n} x_{s\mu} t(x_s \mid \Sigma, x_j)$$

$$\hat{m}_{\mu\nu}{}^2 = \frac{1}{m_0} \frac{1}{n} \sum_{s=1}^{n} (x_{s\mu} - \hat{m}_\mu{}^1)(x_{s\nu} - \hat{m}_\nu{}^1) t(x_s \mid \Sigma, x_j)$$

then

$$\hat{C} = (\hat{m}_1{}^1, \hat{m}_2{}^1, \cdots, \hat{m}_L{}^1)$$

and

$$\hat{R} = \begin{bmatrix} \hat{m}_{11}{}^2 \hat{m}_{12}{}^2 \cdots \\ \hat{m}_{21}{}^2 \hat{m}_{22} \cdots \\ \cdots \quad \hat{m}_{ll}{}^2 \end{bmatrix}. \tag{10}$$

Assuming that the best function is such that relation (9) is true, the above sample moments (10) are used as estimators of the truncated Gaussian function. Thus, estimators of the parameters characterizing that dominating mixture component can be found using the inverse relations (8).

$$\begin{aligned} \hat{\Phi}_d &= (R^{-1} - \Sigma^{-1})^{-1} \\ \hat{\gamma}_d &= (\hat{\Phi}_d \Sigma^{-1} + I)(C - x_j) + x_j. \end{aligned} \tag{11}$$

By (11), a set of parameters $(\hat{\Upsilon}_{d_j}, \hat{\Phi}_{d_j})$ may be associated with x_j, $j = 1, 2, \cdots, n$. (d corresponds to the mixture density belonging to sample x_j.) If the cluster model of separability and $t(x)$ are satisfied and there are a sufficient number of observations, the set $\{(\hat{\Upsilon}_{d_j}, \hat{\Phi}_{d_j})\}$ should be well clustered in the parameter space.

In some special applications, the covariances Φ_i are equal and known such that much of the calculations can be avoided. Then it is sufficient to obtain estimates of just the means of the dominant component:

$$\hat{\gamma}_d = (\Phi \Sigma^{-1} + I)(C - x_j) + x_j \tag{12}$$

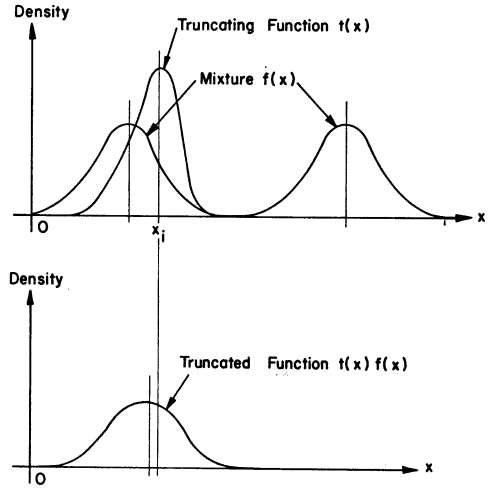

Fig. 3. A graded truncating function.

where $\Phi_i = \Phi$ for all indexes i. If Σ is chosen to be an a^{-1} multiple of Φ, then

$$\hat{\gamma}_d = -ax_j + (1 + a)C. \tag{13}$$

Implementation complexity of this algorithm for the above special case grows with the product of $n^2 l$ instead of the product $n^2 l^2$. In either case, the full data set must be stored.

This approach is especially suited for interactive data analysis and classification where the experimenter supplies the test function covariance matrix Σ from the keyboard. If he assumes that $\Sigma = \Phi_i$, then the simplified expression (13) results. Using a computer output display and assuming a two-dimensional observation space, the experimenter observes that the point x_j maps to the point $\hat{\Upsilon}_d$ according to (13). The effect of mapping all the samples this way is to produce a tighter cluster of points. When dimensionality is greater than two, the problem remains as to how to display the clustered points on a computer output display. Solutions to this problem utilize mappings from a higher dimensional space to either one or two dimensional space such as those in [9], [15], [16].

If $\Phi_i = \Phi$ is unknown, it may not be unreasonable to assume $\Phi = \Sigma$ and make a good initial guess at Φ, say Φ^0 and R,

$$\Phi^n = \frac{1}{n+1} \Phi^0 + \frac{n}{n+1} R$$

where the terms in R are

$$\hat{m}_{\mu\nu}{}^2 = \frac{1}{m_0} \frac{1}{n} \sum_{s=1}^{n} (x_{s_\mu} - \hat{m}_\mu{}^1)(x_{s_\nu} - \hat{m}_\nu{}^1) f(x_s \mid \Phi^{n-1}, x_j)$$

with

$$\hat{m}_0 = \frac{1}{n} \sum_{s=1}^{n} \hat{h}(x_s \mid \Phi^{n-1}, x_j)$$

$$\hat{m}_\mu{}^1 = \frac{1}{\hat{m}_0} \frac{1}{n} \sum_{s=1}^{n} x_{s_\mu} t(x_s \mid \Phi^{n-1}, x_j).$$

Although this estimate of Φ is calculated for the class associated with $\mathbf{x}_j$, it can be used for all classes because of the assumption $\Phi = \Phi_i$. Furthermore, Φ's calculation for various points $\mathbf{x}_j$ can be averaged to produce an estimate of Φ with lower variance.

The conclusion is that the test function approach is a useful technique in an interactive data analysis and classification system when it is used properly within the limitations of the assumptions of the clustering approach. The alternative is a rigorous technique of unsupervised estimation based on a criteria such as $\eta(B)$.

Interactive analysis of data provides the experimenter with results not easily described theoretically. For instance, as shown in the experimental examples in the next section, successive application of this cluster mapping to points mapped from the observation space to the parameter space results in tight clusters in the parameter space.

Experimental Example

Examples using the new clustering algorithm, programmed on the CDC 6500 computer with clusters displayed on the CDC 252 display screen, are presented below; for each example, the test function covariance matrix was determined interactively.

Example 1: The first example is chosen so that there are three well-separated clusters. Specifically, the mixture density (5) is

$$f(x) = \sum_{i=1}^{3} \frac{1}{3} f(\mathbf{x} \mid \Phi_i, \gamma_i) \tag{14}$$

where $f(\mathbf{x} \mid \Phi_i, \Upsilon_i)$ is the bivariate Gaussian distribution with parameters:

$$\Phi_i = \begin{bmatrix} 1.5 & 0 \\ 0 & 1.5 \end{bmatrix}, \qquad i = 1, 2, 3$$
$$\gamma_1 = (0, 0) \tag{15}$$
$$\gamma_2 = (5, -5)$$
$$\gamma_3 = (-5, 5).$$

Two-hundred and fifty observations were independently drawn from a random vector generator designed to have the density (14). These observations are shown in Fig. 4.

The clustering transformation (11) was then used to map the 250 observations to the parameter space of means and covariances. The result was that the observations became more tightly clustered in the parameter space.

The transformation was reapplied to the points in the parameter space which resulted from the transformation of points in the observation space; then the transformation was sequentially applied four times to the points in the parameter space. Fig. 5 shows the results after the fifth application of the clustering transformation. Approximately ten seconds of computation time were required to transform all the observations once.

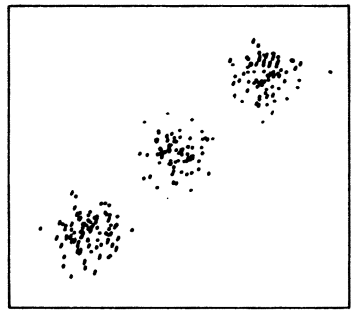

Fig. 4. Original data.

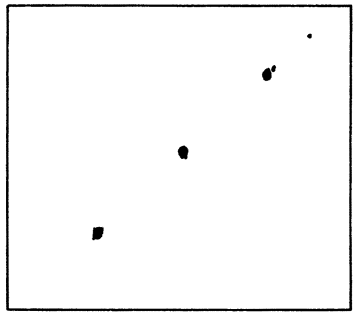

Fig. 5. After fifth mapping.

Fig. 6. Labeled original data.

Example 2: The second example is identical to Example 1 except that the three classes are not as well separated:

$$\Phi_i = \begin{bmatrix} 2.25 & 0 \\ 0 & 2.25 \end{bmatrix}, \qquad i = 1, 2, 3$$
$$\gamma_1 = (4.0, 0) \tag{16}$$
$$\gamma_2 = (-4.0, -2.0)$$
$$\gamma_3 = (4.0, -2.0).$$

One-hundred and fifty observations were drawn randomly with density (14) with parameter values (16). Fig. 6 displays these observations. Fig. 7 indicates the results after the seventh application of the modified cluster algorithm.

Example 3: This third example is a two-class two-dimensional problem where class 1 has a relatively large variance in dimension two, while class 2 has a

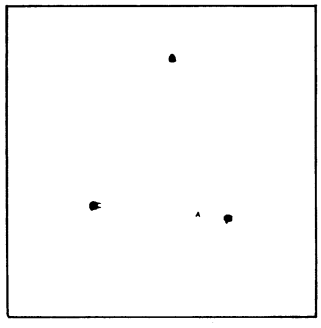

Fig. 7. After seventh mapping

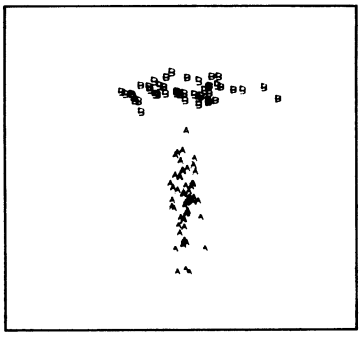

Fig. 8. Original data.

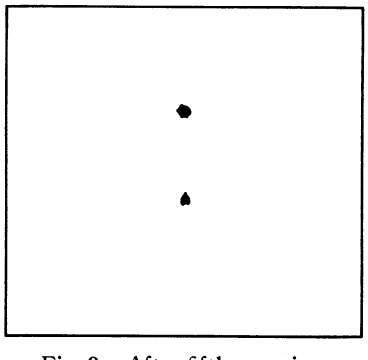

Fig. 9. After fifth mapping.

relatively large variance in dimension one. Specifically, the respective covariance matrices are

$$\Phi_1 = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 4 \end{bmatrix}$$

$$\Phi_2 = \begin{bmatrix} 4 & 0 \\ 0 & \frac{1}{4} \end{bmatrix},$$

the respective mean vectors are

$$\gamma_1 = (0, -2)$$

$$\gamma_2 = (0, 5),$$

and the respective class probabilities are equal.[3] A total of 100 observations were independently drawn from a random vector generator designed to have density (14) for each class. A computer output display of the 100 samples is shown in Fig. 8. *A*'s identify observations from class 1 while *B*'s identify observations from class 2. The mapped samples after five applications of the mapping are shown in Fig. 9.

This algorithm is part of INTERSPACE (Interactive System for Pattern Analysis, Classification, and Enhancement).

References

[1] G. H. Ball, "Data analysis in the social sciences: what about the details," *1965 Fall Joint Computer Conf., AFIPS Proc.*, vol. 27, pt. 1. Washington, D.C.: Spartan, 1965, pp. 533–559.

[2] G. S. Sebestyen, "Pattern recognition by an adaptive process of sample set construction," *IRE Trans. Information Theory*, vol. IT-8, pp. 82–91, September 1962.

[3] R. E. Bonner, "On some clustering techniques," *IBM J. Research and Develop.*, vol. 8, pp. 22–32, January 1964.

[4] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Trans. Royal Society of London*, series A, vol. 185, pp. 77–100, 1894; also, in *K. Pearson, Early Statistical Papers*, reprinted for the Biometrika Trustees, London, England: Cambridge University Press, 1948, pp. 1–40.

[5] C. R. Rao, *Advanced Statistical Methods in Biometric Research.* New York: Wiley, 1952.

[6] D. B. Cooper and P. W. Cooper, "Nonsupervised adaptive signal detection and pattern recognition," *Information and Control*, vol. 7, pp. 416–444, September 1964.

[7] J. C. Hancock and E. A. Patrick, "Learning probability spaces for classification and recognition of patterns with or without supervision," School of Electrical Engineering, Purdue University, Lafayette, Ind., Tech. Rept. 65-21, November 1965.

[8] V. Hasselblad, "Estimation of parameters for a mixture of normal distributions," *Technometrics*, vol. 8, pp. 431–444, August 1966.

[9] E. A. Patrick, D. R. Anderson, and F. K. Bechtel, "Mapping multidimensional space to one dimension for computer output display," *IEEE Trans. Computers*, vol. C-17, pp. 949–953, October 1968.

[10] A. C. Cohen, "Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples," *Ann. Math. Stat.*, vol. 21, pp. 557–569, 1950.

[11] K. S. Miller, *Multidimensional Gaussian Distributions.* New York: Wiley, 1964.

[12] E. A. Patrick and J. P. Costello, "On unsupervised estimation algorithms," *Proc. 1969 IEEE Internatl. Symp. on Information Theory*, January 1969; also, School of Electrical Engineering, Purdue University, Lafayette, Ind., Tech. Rept. 69-18, June 1969.

[13] ——, "On some approaches to unsupervised estimation," School of Electrical Engineering, Purdue University, Lafayette, Ind., Tech. Rept. 68-7, August 1968.

[14] G. Nagy, "State of the art in pattern recognition," *Proc. IEEE*, vol. 56, pp. 836–862, May 1968.

[15] J. W. Sammon, Jr., "A nonlinear mapping for data structure analysis," *IEEE Trans. Computers*, vol. C-18, pp. 401–409, May 1969.

[16] R. N. Shepard and J. D. Carroll, "Parametric representation of nonlinear data structures," in *Multivariate Analysis*, P. R. Kreshnaiak, Ed. New York: Academic Press, 1966.

[3] The two mean vectors and covariance matrices are unknown and estimated according to (12).