A Note on an Adaptive Hypersphere Decision Boundary

PAUL W. COOPER, MEMBER, IEEE

***Abstract*—Supplementing earlier papers, this note treats two topics: *nonsupervised* adaptation to the optimum hypersphere for normal distributions, and *supervised* estimation of the shape parameter *m* for Pearson probability distributions.**

INTRODUCTION

Hyperspherical partitioning of n-space into two category regions involves a simple decision procedure wherein an unknown $\mathbf{x}$ is assigned to one category if $(\mathbf{x}-\mathbf{c})'(\mathbf{x}-\mathbf{c}) \leq R_n^2$, otherwise to the remaining one, where $\mathbf{c}$ and R_n are, respectively, the sphere center and radius.[1] In various *pattern recognition* problems and in certain problems in signal detection[2] [4] the probability distributions are spherically symmetric with different variances (and different means), and a spherical decision rule is optimum if these distributions are normal or Pearson Types II or VII [1]. As a supplement to that earlier paper [1], the end of this note contains relationships for estimating the distribution type and the shape parameter for Pearson distributions.

In another work [3] it was shown that it is possible in many cases to achieve *nonsupervised* adaptation wherein the desired descision parameters are estimated from a sequence of "learning" samples whose category associations are *unknown* a priori. The procedures included there for linear decision rules are of simple form. *Nonsupervised* adaptation can also be achieved simply for the hypersphere partition, as illustrated here for two spherically symmetric multivariate normal distributions (Fig. 1)

$$P_k(\mathbf{x}) = (2\pi)^{-n/2}\sigma_k^{-n} \exp\left[-(1/2\sigma_k^2)(\mathbf{x}-\mathbf{u}_k)'(\mathbf{x}-\mathbf{u}_k)\right], \qquad k = 1, 2.$$

A NONSUPERVISED ADAPTIVE HYPERSPHERE FOR NORMAL DISTRIBUTIONS

Associating a priori probability q with category 1, the parameters necessary for defining the partitioning hypersphere are $\mathbf{u}_1$, $\mathbf{u}_2$, σ_1, σ_2, and perhaps q. Approaching the problem of *nonsupervised* estimation by viewing the samples as being drawn from an over-all distribution comprised of the two category distributions, in this section the parameters listed are for the over-all distribution, unless specifically indexed with a subscript referring to a category. Inspired by the desire to have convergent estimators which are simple to implement, we have used *moments* of the over-all distribution for determining the parameters (of the component distributions) in terms of which the decision rule is defined.

Same Means

The mean $\boldsymbol{\mu} = E(\mathbf{x})$ and the νth central radial moment,

$$\psi_\nu = E([(\mathbf{x}-\boldsymbol{\mu})'(\mathbf{x}-\boldsymbol{\mu})]^{\nu/2}), \tag{1}$$

are estimated from M "learning" samples $\{\mathbf{x}_j\}$, $j=1, 2, \cdots, M$. Sample moments can be used, e.g.,

$$\hat{\psi}_\nu = \frac{1}{M-1}\sum_{j=1}^{M}[(\mathbf{x}_j-\hat{\boldsymbol{\mu}})'(\mathbf{x}_j-\hat{\boldsymbol{\mu}})]^{\nu/2}.$$

From now on we discuss relations (for normal distributions) in terms of true moments. In practice the estimated moment values would be substituted in the equations which follow. Two central radial moments in n-space are:

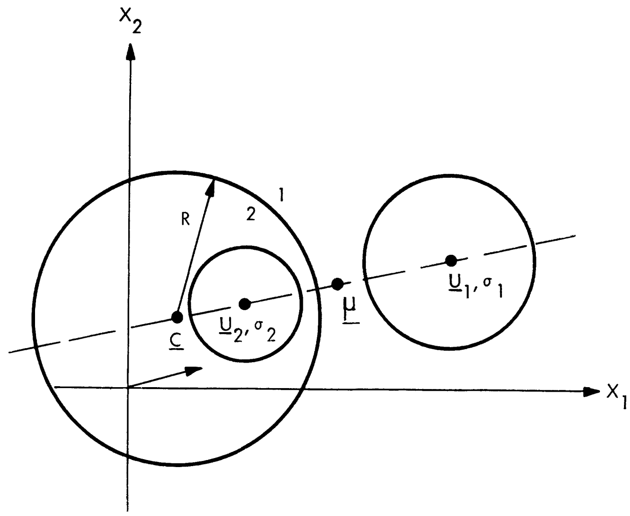

Fig. 1. Spherical decision rule in 2-space.

$$\begin{aligned}\psi_2 &= [q\sigma_1^2 + (1-q)\sigma_2^2]n\\ \psi_4 &= [q\sigma_1^4 + (1-q)\sigma_2^4]n(n+2).\end{aligned} \tag{2}$$

If it is not zero, $\boldsymbol{\mu}$ is directly estimated. For q known (where we have associated $1-q$ with category 2, which is treated as the one with the tighter distribution), the unknown σ_1 and σ_2 are determined from ψ_2 and ψ_4. (They could of course be determined from any pair of moments.)

$$\begin{aligned}\sigma_2^2 &= \frac{\psi_2}{n} - \frac{1}{n}\left(\frac{q}{1-q}\right)^{1/2}\left(\frac{n}{(n+2)}\psi_4 - \psi_2^2\right)^{1/2}\\ \sigma_1^2 &= \frac{\psi_2}{n} + \frac{1}{n}\left(\frac{1-q}{q}\right)^{1/2}\left(\frac{n}{(n+2)}\psi_4 - \psi_2^2\right)^{1/2}.\end{aligned} \tag{3}$$

(Note, for q much different from $\frac{1}{2}$ the estimate of the variance of the lower probability distribution, as in (3), will be sensitive to estimation error in the moments.)

For $q=(1-q)=\frac{1}{2}$ these expressions simplify, as they do for $n=1$. One could work instead with the marginal moments, corresponding to $n=1$, perhaps first averaging the marginal sample moments in the n coordinate directions.

For σ_2 known and σ_1 and q unknown, corresponding to detection of unknown signal of unknown probability q in known noise, σ_1 and q are determined:

$$(1-q) = \frac{\psi_2^2/n^2 - \psi_4/n(n+2)}{2\psi_2\sigma_2^2/n - \psi_4/n(n+2) - \sigma_2^4}, \tag{4}$$

which is then substituted in

$$\sigma_1^2 = (1/q)[\psi_2/n - (1-q)\sigma_2^2]. \tag{5}$$

Different Means

When the two category distributions have different means, the over-all distribution can be viewed as being "bimodal," (where strictly speaking true bimodality does not become apparent unless the component distributions are sufficiently displaced from one another). First consider equal a priori probabilities. Central to this discussion are the mean $\boldsymbol{\mu}$ and the over-all covariance matrix $\mathbf{B}=[b_{ij}]$, where

$$\boldsymbol{\mu} = \tfrac{1}{2}(\mathbf{u}_1+\mathbf{u}_2), \qquad \mathbf{B} = \boldsymbol{\alpha}\boldsymbol{\alpha}' + \tfrac{1}{2}\mathbf{I}(\sigma_1^2+\sigma_2^2) \tag{6}$$

and

$$\boldsymbol{\alpha} = (1/2)(\mathbf{u}_2-\mathbf{u}_1) = \mathbf{u}_2 - \boldsymbol{\mu} = \boldsymbol{\mu} - \mathbf{u}_1.$$

Manuscript received May 12, 1966. This work was supported under Contract AF30(602)-3619 by Rome Air Development Center, Griffiss AFB, Rome, N. Y.

The author is a consultant, located at 2 Forest Street, Cambridge, Mass. He was formerly with the Applied Research Laboratory, Sylvania Electronics Systems, Waltham, Mass.

[1] Boldface denotes a (column) vector or a matrix in n-space, and a prime its transpose.

[2] For example, in detection of a sampled (band-limited white) Gaussian signal in similar type noise. The hypersphere partition further arises in situations involving a flat-fading Rician channel, for reception of two constant signals, or for detection of the presence of a signal having component samples of equal magnitude. For the received $\mathbf{n}$-vector $\mathbf{x}$, and representing the additive noise by $\mathbf{n}$, the transmitted signal by $\mathbf{s}$, and the flat-fading effect by the diagonal matrix $\mathbf{F}$ having random elements which are independent and zero-mean normal with equal variance, and $\mathbf{I}$ being the identity matrix, we have the following relation for the flat-fading Rician channel: $\mathbf{x}=\mathbf{n}+(\mathbf{I}+\mathbf{F})\mathbf{s}$.

By any one of several methods, we obtain the principal direction of the "bimodal" distribution, corresponding to the major eigenvector of $\boldsymbol{B}$.

The smaller eigenvalue λ_2, having multiplicity $(n-1)$, and the principal one λ_1 are

$$\lambda_2 = \tfrac{1}{2}(\sigma_1^2 + \sigma_2^2), \qquad \lambda_1 = \boldsymbol{\alpha}'\boldsymbol{\alpha} + \lambda_2. \tag{7}$$

There are several avenues for achieving solution. One is to draw upon equations (21)–(28) of ref. [3] to obtain λ_2 in terms of the trace and the sum of the 2×2 determinants on the diagonal of $\boldsymbol{B}$, obtaining then $\boldsymbol{\alpha}$ from $\alpha_i^2 = b_{ii} - \lambda_2$ and from the off-diagonal elements of $\boldsymbol{B}$ (taking α_1 positive and α_i the same sign as b_{i1}). Then make use of Φ_p the third central moment in the principal direction. Take $\boldsymbol{e}_p$ as a unit vector in the principal direction, scalar α as $\boldsymbol{\alpha}'\boldsymbol{e}_p$, and $\Phi_p = E([(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{e}_p]^3)$,

$$\Phi_p = -\tfrac{3}{2}\alpha(\sigma_1^2 - \sigma_2^2) \tag{8}$$

to obtain

$$\sigma_2^2 = \lambda_2 + \frac{\Phi_p}{3\alpha}, \qquad \sigma_1^2 = \lambda_2 - \frac{\Phi_p}{3\alpha}. \tag{9}$$

Note, the significance of the subscripts 1 and 2 in this different-means treatment is to maintain the correct associations among the σ's and α, and subsequently with q. After solving for the parameters, if we still want to maintain the convention for the decision procedure of calling category 2 the tighter one, if necessary reverse the indexing of the subscripts. For example, in (9), for α as defined earlier, if $\sigma_2 < \sigma_1$, Φ_p will have opposite algebraic sign from α.

Another avenue is to make use of equations (29)–(30) of ref. [3] wherein $\alpha_1^2 = b_{21}b_{31}/b_{32}$, taking the positive root, and $\alpha_i = b_{i1}/\alpha_1$. λ_2 is obtained either as above or from the second moment in the $(n-1)$-space complementary to the direction of $\boldsymbol{\alpha}$, and again (9) is used to obtain σ_1 and σ_2.

A third avenue is to make use of the iterative techniques (considered in the earlier paper) for obtaining the principal eigenvector. Both λ_1 and λ_2 can be obtained iteratively, or λ_2 can be found as indicated above with equations (21)–(26). Then (using symbols of those equations)

$$\lambda_1 - \lambda_2 = T_{n-1} - n\lambda_2 = \alpha^2, \tag{10}$$

and (9) is used for σ_1 and σ_2.

It is worthwhile to note that once the principal direction has been obtained, *the remaining $(n-1)$-space can be treated as the case of the same means*, since $\boldsymbol{u}_2 - (\boldsymbol{u}_2'\boldsymbol{e}_p)\boldsymbol{e}_p = \boldsymbol{u}_1 - (\boldsymbol{u}_1'\boldsymbol{e}_p)\boldsymbol{e}_p$, and the relations for the equal-means case can then be applied directly.

When some parameters are known there is again much simplification. For example, for one category mean known, the other is obtained from the overall mean $\boldsymbol{\mu}$. Then two moments are used to obtain the variances: use the second moment λ_2 and the third Φ_p in the principal direction as in (9), or use the second and fourth moments in the remaining $(n-1)$-space invoking the common means relation of (3) for $q = \frac{1}{2}$ and replacing n with $(n-1)$.

For known $q \neq \frac{1}{2}$, the previous relations are modified somewhat. Taking $\boldsymbol{u}_2 - \boldsymbol{\mu} = \boldsymbol{t}$ and $\boldsymbol{u}_1 - \boldsymbol{\mu} = \boldsymbol{s}$, and scalar $t = \boldsymbol{t}'\boldsymbol{e}_p$ and $s = \boldsymbol{s}'\boldsymbol{e}_p$,

$$\begin{aligned}
\boldsymbol{\mu} &= q\boldsymbol{u}_1 + (1-q)\boldsymbol{u}_2 \\
\boldsymbol{s} &= -\left(\frac{1-q}{q}\right)\boldsymbol{t} \\
\boldsymbol{B} &= q\boldsymbol{s}\boldsymbol{s}' + (1-q)\boldsymbol{t}\boldsymbol{t}' + I[q\sigma_1^2 + (1-q)\sigma_2^2] \\
&= \left(\frac{1-q}{q}\right)\boldsymbol{t}\boldsymbol{t}' + I[q\sigma_1^2 + (1-q)\sigma_2^2] \\
\Phi_p &= -3t(1-q)(\sigma_1^2 - \sigma_2^2) - t^3\left(\frac{1-q}{q}\right)\left(\frac{1-2q}{q}\right) \\
\lambda_1 - \lambda_2 &= \left(\frac{1-q}{q}\right)\boldsymbol{t}'\boldsymbol{t}
\end{aligned} \tag{11}$$

and

$$\begin{aligned}
\sigma_2^2 &= \lambda_2 + \frac{\Phi_p}{3t}\left(\frac{q}{1-q}\right) + \frac{t^2}{3}\left(\frac{1-2q}{q}\right) \\
\sigma_1^2 &= \lambda_2 + \frac{\Phi_p}{3s}\left(\frac{1-q}{q}\right) - \frac{s^2}{3}\left(\frac{1-2q}{1-q}\right) \\
&= \lambda_2 - \frac{\Phi_p}{3t} - \frac{t^2}{3}\left(\frac{1-q}{q}\right)\left(\frac{1-2q}{q}\right).
\end{aligned} \tag{12}$$

As before, a priori knowledge of one or more of the parameters leads to simplified expressions, e.g., if σ_2 is known but q is not, then the common-means solution can be used in the $(n-1)$-space complementary to the principal direction.

Supplement to "The Hypersphere in Pattern Recognition"

In reference [1], where it was shown that the hypersphere partition is optimum for Pearson Types II and VII distributions as well as for the normal, *supervised* estimators were presented for the means $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ and the scaling parameters ω_1 and ω_2, each of the latter two inversely proportional to its associated standard deviation σ_1 and σ_2 (where it exists). Supervised estimation of q, where needed, is obvious. The previously presented estimators for the scale parameters and the partitioning hypersphere parameters themselves depend also upon the distribution type and upon the important distribution shape parameter m, relationships for which are presented in this section.

Maximum-likelihood estimators are easy to formulate but for these distributions are algebraically formidable. Sample moments are easy to work with and are good (except for Pearson Type VII with low value for m). Equations (25), (26), (29), (30), and (34) of [1] define the Pearson distributions and give some of the estimators.

Treating a single distribution, in terms of (1) we define a kurtosis in n-space as

$$\beta_2 = \psi_4/\psi_2^2, \tag{13}$$

noting that this β_2 differs from the β_2 of Pearson [5] unless n is taken to be 1. β_2 is estimated in terms of sample moments.

For the normal distribution, which is transitional between the Pearson Types II and VII,

$$\beta_2 = 1 + \frac{2}{n}. \tag{14}$$

Types II and VII have, respectively, smaller and larger β_2. Therefore, examination of β_2 can be used to determine distribution type.

The shape parameter m for the multivariate distribution is

$$m = \frac{\beta_2(n/2+2) - (n/2+2+2/n)}{|1 + 2/n - \beta_2|} \tag{15}$$

If desired, one could proceed instead by using an estimate of the univariate marginal kurtosis to obtain m by first obtaining the shape parameter of the marginal distribution [2], [5].

For low m for Type VII the integer moments of interest may not exist, and therefore β_2 may be inapplicable. Other types of estimators which are suitable and tractable involve use of order statistics or of fractional moments (reference [2], pp. 129–133).

Acknowledgment

I should like to thank Miss Ellen Haczynski for typing this short manuscript with excellence, as she has many longer ones in the past.

References

[1] P. W. Cooper, "The hypersphere in pattern recognition," *Information and Control*, vol. 5, pp. 324–346, December 1962.

[2] ——, "Hyperplanes, hyperspheres, and hyperquadrics as decision boundaries," *Computer and Information Sciences*, Washington, D.C.: Spartan, 1964, ch. 4, pp. 111–138.

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

[4] D. Middleton, "On the detection of stochastic signals in additive normal noise," *IRE Trans. on Information Theory*, vol. IT-3, pp. 86–121, June 1957.

[5] K. Pearson, "Contributions to the mathematical theory of evolution, II, Skew variation in homogeneous material," *Philosophical Transactions A*, vol. 186, pp. 343–414, 1895.