1. Introduction

The author previously treated the problem of classification in discrete cases, employing the notion of distance [1]. The purpose of this paper is to treat that problem for multivariate Gaussian cases from the same point of view.

Now, the classification problem is formulated as follows. Let \( \{ \omega_i \} \) be a class of sets of distributions, and let \( X \) be a random variable under consideration. Then the problem is to decide which \( \omega_i \) is considered to contain the distribution of \( X \). We, of course, assume here that \( \omega_i \) and \( \omega_j \) have no common distributions when \( i \neq j \). Further, for efficient decision making we assume that for a suitable distance \( d(\cdot, \cdot) \) in the space of distributions concerned, we have \( d(\omega_i, \omega_j) > \alpha \) (> 0), \( (i \neq j) \). In some cases, when \( d(\omega_i, \omega_j) = 0 \), we can represent each of those \( \omega_i \) by a single distribution \( F_i \), so that \( d(F_i, F_j) > 0 \). For such \( F_i \), we can consider the averaged distribution of \( \omega_i \) by an adequate distribution over \( \omega_i \).

When the distributions concerned are all known, the decision rule for the above problem runs as follows. Let \( S_n \) be an 'empirical' distribution based on \( n \) observations on \( X \). Suppose that \( X \) has one of \( F_1, \ldots, F_t \) as its distribution. Let \( S_n, S_{n,n} \) denote the 'empirical' distribution based on a sample of \( n \) from \( F_i \) which has the same form as \( S_n \). Then we consider \( d(S_n, S_{n,n}) \) and take \( F_i \) when \( S_n, S_{n,n} \) minimizes \( d(S_n, S_{n,n}) \).
Since the case of a finite number of distributions can reduce to the case where
the number of distributions concerned is 2, we shall confine our consideration
to this case.

Let \( F_1, F_2 \) be the distributions concerned, and let \( S'_1, S''_1 \) be the ‘empirical’
distributions determined by observations on \( F_1 \) and \( F_2 \), respectively. Then the
decision rule for this case is the following:

(i) when \( d(S_n, S'_1) < d(S_n, S''_1) \), we decide on \( F_1 \);

and

(ii) when \( d(S_n, S'_1) > d(S_n, S''_1) \), we decide on \( F_2 \).

(iii) For the case \( d(S_n, S'_1) = d(S_n, S''_1) \), we determine in advance to take
either of \( F_1, F_2 \), say \( F_1 \).

The success rate is given by

\[
P(d(S_n, S'_1) \leq d(S_n, S''_1) | F_1),
\]

where “\( |F_1 \)” in the parentheses means “under the condition that \( X \) has \( F_1 \)” and

\[
P(d(S_n, S'_1) > d(S_n, S''_1) | F_2).
\]

Now, when \( d(\cdot, \cdot) \) satisfies the triangle axiom, we obtain

\[
d(S_n, S'_1) \leq d(S_n, F_1) + d(F_1, S'_1),
\]

\[
d(S_n, S''_1) \geq d(F_1, F_2) - d(F_1, S_n) - d(F_2, S''_1),
\]

and

\[
d(S_n, S'_1) - d(S_n, S''_1) \geq d(F_1, F_2) - 2d(S_n, F_1) - d(F_1, S'_1) - d(F_2, S''_1).
\]

Therefore, when \( d(F_1, F_2) \geq \delta (> 0) \), we have

\[
d(S_n, S'_1) - d(S_n, S''_1) \geq \delta - 2d(S_n, F_1) - d(F_1, S'_1) - d(F_2, S''_1),
\]

and further, when \( 2d(S_n, F_1) + d(F_1, S'_1) + d(F_2, S''_1) \leq \delta \), we have \( d(S_n, S'_1) \leq d(S_n, S''_1) \). As a result we obtain

\[
P(d(S_n, S'_1) \leq d(S_n, S''_1) | F_1)
\]

\[
\geq P(2d(S_n, F_1) + d(F_1, S'_1) + d(F_2, S''_1) < \delta | F_1)
\]

\[
\geq P \left( d(S_n, F_1) < \frac{\delta}{4}, d(F_1, S'_1) < \frac{\delta}{4}, d(F_2, S''_1) < \frac{\delta}{4} \right | F_1 \)
\]

\[
= P \left( d(S_n, F_1) < \frac{\delta}{4} \right | F_1 \) \cdot P \left( d(F_1, S'_1) < \frac{\delta}{4} \right | F_1 \) \cdot P \left( d(F_2, S''_1) < \frac{\delta}{4} \right).
\]

Thus, for evaluation of the success rate it is sufficient to know about

\[
P \left( d(S_n, F) < \frac{\delta}{4} \right).
\]

3. Distance and ‘test’ statistics

Let \( F_1, F_2 \) be distributions defined in space \( R \), and let \( p_1(x), p_2(x) \) be their
density functions with respect to a measure \( m \) in \( R \). Then the distance between
distributions which we employ here is
This distance satisfies the metric space axioms. When we define
\[
d(\mathcal{F}_1, \mathcal{F}_2) = \left[ \int_{\mathbb{R}} (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 \, dm \right]^{1/2},
\]
we have
\[
\rho(\mathcal{F}_1, \mathcal{F}_2) = \int_{\mathbb{R}} \sqrt{p_1(x)} \, \sqrt{p_2(x)} \, dm,
\]
(11)

The quantity \( \rho(\cdot, \cdot) \) expresses the closeness between distributions, and we can use \( \rho(\cdot, \cdot) \) in place of \( d(\cdot, \cdot) \).

Now, let us turn to the multivariate Gaussian case.

Let \( \mathcal{R} \) be a \( k \)-dimensional space, and let \( \mathcal{F}_1, \mathcal{F}_2 \) be \( k \)-dimensional Gaussian distributions with density functions
\[
p_1(x) = \frac{|A|^{1/2}}{(2\pi)^{k/2}} \exp \left[ -\frac{1}{2} (x - a) (x - a)' A^{-1} \right],
\]
(12)
\[
p_2(x) = \frac{|B|^{1/2}}{(2\pi)^{k/2}} \exp \left[ -\frac{1}{2} (x - b) (x - b)' B^{-1} \right],
\]
(13)
where \( A, B \) are positive-definite matrices of degree \( k \), and \( x, a, b \) are \( k \)-dimensional (column) vectors. Then we obtain
\[
\rho(\mathcal{F}_1, \mathcal{F}_2) = \frac{|AB|^{1/4}}{|\frac{1}{2} (A + B)|^{1/2}} \exp \left[ -\frac{1}{8} \left\{ (A + B)^{-1} (Aa + Bb), (Aa + Bb) \right\} \right.
\]
\[
+ (Aa, a) + (Bb, b) \}
\]
(see [2]). When \( A = B \),
\[
\rho(\mathcal{F}_1, \mathcal{F}_2) = \exp \left[ -\frac{1}{8} (a - b) (a - b) \right].
\]
When \( a = b \),
\[
\rho(\mathcal{F}_1, \mathcal{F}_2) = \frac{|AB|^{1/4}}{|\frac{1}{2} (A + B)|^{1/2}}.
\]

Let \( X_1, X_2, \ldots, X_n \) be \( n \) \((\geq 2)\) observations on a random variable \( X \) with a \( k \)-dimensional Gaussian distribution. Define
\[
\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,
\]
(17)
\[
V = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})',
\]
(18)
and let \( S_n \) be the \( k \)-dimensional Gaussian distribution with mean \( \bar{X} \) and covariance matrix \( V \), that is, \( S_n = N(\bar{X}, V) \). Set \( U = V^{-1} \). Similarly, concerning \( F_1 \) and \( F_2 \), let
\[
S'_n = N(\bar{X}_n, V_n), \quad U_n = V_n^{1/2},
\]
(19)
\[
S''_n = N(\bar{X}_n, V_n), \quad U_n = V_n^{1/2}.
\]
(20)
Then we have

\[ p(S_n, S'_1) = \frac{|UU_1|^{1/4}}{\left(\frac{1}{2}(U + U_1)\right)^{1/2}} \exp \left\{ -\frac{1}{4} \left\{ -(U + U_1)^{-1}(U\overline{X} + U\overline{X}_1), (U\overline{X} + U\overline{X}_1) + (U\overline{X}, \overline{X}) + (U\overline{X}_1, \overline{X}_1) \right\} \right\}, \]

\[ p(S_n, S'_2) = \frac{|UU_2|^{1/4}}{\left(\frac{1}{2}(U + U_2)\right)^{1/2}} \exp \left\{ -\frac{1}{4} \left\{ -(U + U_2)^{-1}(U\overline{X} + U\overline{X}_2), (U\overline{X} + U\overline{X}_2) + (U\overline{X}, \overline{X}) + (U\overline{X}_2, \overline{X}_2) \right\} \right\}. \]

Using these statistics we can make a decision; that is, when \( p(S_n, S'_1) \geq p(S_n, S'_2) \), we decide that \( X \) has \( F_1 \), and when \( p(S_n, S'_1) < p(S_n, S'_2) \), we decide that \( X \) has \( F_2 \). When it is known in advance that \( A = B \), we consider

\[ p_1(S_n, S'_1) = \exp \left\{ -\frac{1}{4}(U\overline{X} - \overline{X}_1), (\overline{X} - \overline{X}_1) \right\}, \]

\[ p_1(S_n, S'_2) = \exp \left\{ -\frac{1}{4}(U\overline{X} - \overline{X}_2), (\overline{X} - \overline{X}_2) \right\} \]

for the case where \( A = B \) is unknown, and

\[ p_2(S_n, S'_1) = \exp \left\{ -\frac{1}{4}(A\overline{X} - \overline{X}_1), (A\overline{X} - \overline{X}_1) \right\}, \]

\[ p_2(S_n, S'_2) = \exp \left\{ -\frac{1}{4}(A\overline{X} - \overline{X}_2), (A\overline{X} - \overline{X}_2) \right\} \]

for the case where \( A = B \) is known.

When the problem is concerned only with the covariance matrix, we consider

\[ p_3(S_n, S'_1) = \frac{|UU_1|^{1/4}}{\left(\frac{1}{2}(U + U_1)\right)^{1/2}}, \]

\[ p_3(S_n, S'_2) = \frac{|UU_2|^{1/4}}{\left(\frac{1}{2}(U + U_2)\right)^{1/2}}. \]

For instance, when

\[ \frac{|UU_1|^{1/4}}{\left(\frac{1}{2}(U + U_1)\right)^{1/2}} \geq \frac{|UU_2|^{1/4}}{\left(\frac{1}{2}(U + U_2)\right)^{1/2}}, \]

we decide that \( X \) has \( F_1 \), and when

\[ \frac{|UU_1|^{1/4}}{\left(\frac{1}{2}(U + U_1)\right)^{1/2}} < \frac{|UU_2|^{1/4}}{\left(\frac{1}{2}(U + U_2)\right)^{1/2}}, \]

we decide that \( X \) has \( F_2 \). (For the case where these two statistics are equal, we can, of course, determine in advance to take \( F_2 \).)

As to the success rate, we obtain

\[ P \left( p(S_n, S'_1) > p(S_n, S'_2) \middle| F_1 \right) \geq P \left( p(S_n, F_1) > \frac{1 - \delta}{16} \right) \]

\[ \times P \left( p(S'_1, F_1) > \frac{1 - \delta}{16} \right) \cdot P \left( p(S'_2, F_2) > \frac{1 - \delta}{16} \right), \]

and from this relation we can get an evaluation of the success rate, when we
have the value of $P(p(F, S_n) > \delta|F)$. Thus the next problem is to evaluate $P(p(F, S_n) > \delta|F)$ ($\delta < 1$).

Assume that $X$ is distributed according to $N(a, \Sigma)$. First, concerning $p_1(F, S_n)$, $p_2(F, S_n)$, we have

(32) \[-8n \log p_1(F, S_n) = n(V^{-1}(X - a), (X - a)),\]

(33) \[-8n \log p_2(F, S_n) = n(\Sigma^{-1}(X - a), (X - a)),\]

and, as is well known, the right-hand sides have a noncentral $F$ and a chi-square distribution, and we have no problem here.

Concerning $p_3(F, S_n)$, we have

(34) \[
p_3(F, S_n) = \frac{|\Sigma^{-1}U|^{1/4}}{|\frac{1}{2}(\Sigma^{-1} + U)|^{1/2}},
\]

and

(35) \[
P(p_3(F, S_n) > \delta) \geq \left[ P\left( \frac{4Z}{(1 + Z)^2} > \delta \right) \right]^k,
\]

where $Z$ is a random variable such that $nZ$ has the chi-square distribution with $n$ degrees of freedom (see [2]). Therefore, for given positive $\delta$ and $\epsilon$ ($\epsilon < 1$), there exists an integer $n_0$ such that $P(p_3(F, S_n) > \delta) \geq 1 - \epsilon$ uniformly in $F$ for $n \geq n_0$.

Now we will present the general case. Let $\delta_1$, $\delta_2$ be positive numbers such that $\delta = \delta_1 \exp \left[-\frac{1}{4}\delta_2\right]$, $\delta_1 < 1$. Then we get

(36) \[
P(p(F, S_n) > \delta) \geq P\left( \frac{|\Sigma^{-1}W^{-1/4}}{|\frac{1}{2}(\Sigma^{-1} + W^{-1})|^{1/2}} > \delta_1 \right) P(\beta(\Sigma^{-1}(X - a), (X - a)) < 2\delta_2)
\]

where

(37) \[
W = \begin{pmatrix}
\frac{1}{n} \sum_{i=1}^{n} X_{i1}^2 & 0 \\
0 & \frac{1}{n} \sum_{i=1}^{n} X_{ik}^2
\end{pmatrix},
\]

(38) \[
\beta = 2\frac{\delta_1}{\delta_1^2} - 1 + \frac{2}{\delta_1^2} \sqrt{1 - \delta_1^4}
\]

(see [2]). By taking $\delta_1$ (accordingly $\delta_2$) so that the right-hand side becomes maximum, we can get an evaluation (from below) of $P(p(F, S_n) > \delta|F)$. When we want to have $P(p(F, S_n) > \delta) > 1 - \epsilon$, let $1 - \epsilon = \alpha_1\alpha_2$, $\alpha_1$, $\alpha_2 > 0$ and take $n$ large so that

(39) \[
P\left( \frac{|\Sigma^{-1}W^{-1/4}}{|\frac{1}{2}(\Sigma^{-1} + W^{-1})|^{1/2}} > \delta_1 \right) \geq \alpha_1,
\]

(40) \[
P(\beta(\Sigma^{-1}(X - a), (X - a)) < 2\delta_2) \geq \alpha_2.
\]
4. Classification by a linear function of vector components

In this section we consider the classification problem by a linear function of components of a random vector. Let \( N(a^{(1)}, \Sigma_1), N(a^{(2)}, \Sigma_2) \) be \( k \)-dimensional Gaussian distributions, and let \( X = (X_1, \cdots, X_k) \) be a \( k \)-dimensional random vector. The problem is to decide which one of \( N(a^{(1)}, \Sigma_1), N(a^{(2)}, \Sigma_2) \) is the distribution of \( X \). For this problem we consider a linear function of the components of \( X \) of the form \( (c, X) = c_1 X_1 + \cdots + c_k X_k \), where \( c \) is a constant vector \( (\neq 0) \). The decision procedure is as follows. Let \( X^{(1)}, X^{(2)} \) be samples from \( N(a^{(1)}, \Sigma_1), N(a^{(2)}, \Sigma_2) \), and let \( F_{1c}, F_{2c} \) be the distributions of \( (c, X^{(1)}) \) and \( (c, X^{(2)}) \). Further, let \( E_1, E_2 \) be optimal regions (on the real line) for classifying an observation from \( F_{1c} \) or \( F_{2c} \). (For instance, \( E_1, E_2 \) can be defined by the probability ratio rule.) Then, when \( (c, X) \) lies in \( E_1 \), we decide that \( X \) has \( N(a^{(1)}, \Sigma_1) \), and when \( (c, X) \) lies in \( E_2 \), we decide that \( X \) has \( N(a^{(2)}, \Sigma_2) \). Therefore, for reducing the probability of misclassification, it is necessary to find an adequate \( c \).

Now, we have

\[
E(c, X^{(1)}) = (c, a^{(1)}), \\
V(c, X^{(1)}) = (c, \Sigma_{1c}), \\
E(c, X^{(2)}) = (c, a^{(2)}), \\
V(c, X^{(2)}) = (c, \Sigma_{2c}),
\]

and

\[
\rho(F_{1c}, F_{2c}) = \left[ \frac{2(c, \Sigma_{1c})^{1/2}(c, \Sigma_{2c})^{1/2}}{(c, \Sigma_{1c}) + (c, \Sigma_{2c})} \right]^{1/2} \exp \left[ -\frac{1}{4} \frac{(c, a^{(1)} - a^{(2)})^2}{(c, \Sigma_{1c})} \right].
\]

Therefore, from our standpoint, we should choose a \( c \) that minimizes \( \rho(F_{1c}, F_{2c}) \) when \( a^{(1)}, a^{(2)}, \Sigma_1, \Sigma_2 \) are known. When \( a^{(1)}, a^{(2)}, \Sigma_1, \Sigma_2 \) are unknown, we use in place of them their estimates obtained from samples.

For example, when it is known beforehand that \( \Sigma_1 = \Sigma_2 \), we consider

\[
\frac{(c, a^{(1)} - a^{(2)})^2}{(c, \Sigma_{1c})}
\]

and determine \( c \) so as to maximize this value. (This is a familiar procedure in multivariate analysis.)

REFERENCES