Res. Rep. Fac. Eng. Mie Univ., Vol. 23, pp. 51−70 (1998)

Original Paper

Limited Sample Based Optimum Classifier Design and the Evaluation of the Mean Error Rate

Fumitaka KIMURA, Tetsushi WAKABAYASHI, Yasuji MIYAKE

(Received September 14, 1998)

Abstract

This paper deals with limited sample based optimum classifier design and the theoretical evaluation of the mean error rate. Gaussian population with unknown parameters is assumed. The conditional density given a limited sample of the population is first derived, and its relationship to the multivariate t-distribution is shown. Then, the mean error rate of the optimum classifier is theoretically evaluated by the quadrature of the conditional density. To verify the optimality of the classifier and the correctness of the mean error calculation, the results of Monte Carlo simulation employing a new sampling procedure are shown. The role of the a priori distribution in reducing the mean error rate is discussed at the end of this paper.

Key words: Statistical pattern recognition, optimum classifier, Bayes error, limited sample effect.

1 Introduction

In order to design robust pattern classifiers for real world applications, a limited sample theory of pattern recognition is expected to be of essential importance. As a part of the theory, this paper deals with limited sample based optimum classifier design and the theoretical evaluation of the mean error rate. Gaussian population with unknown parameters is assumed. To derive the optimum discriminant function, the conditional density given a limited sample of the population is first derived, and its relationship to the multivariate t-distribution is shown. As a result, the obtained optimum classifier is different from the conventional quadratic classifier known to be optimum for Gaussian distributions with known parameters. Especially when the sample size of classes are not equal, the optimum discriminant function is not quadratic, and the decision surface is not hyperquadratics.

Then, the mean error rate of the optimum classifier is theoretically evaluated by the quadrature of the conditional density. For univariate case, the mean error rate of two-class problem with different sample size and different sample covariance matrixes is evaluated. For multivariate case, the one with common sample size, common sample covariance matrixes, and common *a priori* probabilities is evaluated. Since these mean error rates are obtained by taking the expectation of the error rate over unknown population parameters dealt as random variables, they only depend on known parameters such as sample parameters, sample size, and the dimensionality. In this point, the presented mean error rate has its own interpretation and significance different from those of conventional mean error rate which requires the unknown population parameters in its calculation. To verify the optimality of the classifier and the correctness of the mean error calculation, the results of Monte Carlo simulation employing a new sampling procedure are shown.

This paper is based on the Bayesian classic theory which deals with unknown parameters as random variables and assumes their *a priori* distributions. The essential of the *a priori* distribution has not been completely known yet, and the validity of the Bayesian approach and its application has been long argued [1, p. 76]. The fact that the Bayesian approach enables us to design the optimum classifier based on limited sample and to evaluate the mean error rate using known parameters alone clearly demonstrates the validity of this approach. The role of the *a priori* distribution in reducing the mean error rate is discussed at the end of this paper.

The optimum classifier based on a limited sample was first derived by Keehn [2]. He studied the asymptotic properties of the optimum classifier and calculated type I error, which is the rejection rate for a given threshold value of the likelihood. However the mean error rate for two-class problem was not evaluated, and the properties of the optimum classifier except for the asymptotic properties were not studied. This paper also shows minor modification and correction required in his derivation. Although

the approximate relationship between the conditional density and the multivariate t-distribution was pointed out in [3], the exact relationship has not been established.

The mean error rate of two-class Gaussian problem with unknown mean vectors and unknown covariance matrixes was evaluated by Sitgreaves [4]. However her formula contains five fold infinite arithmetic series and requires the unknown population parameters in its calculation, which make the formula theoretically and practically intractable. Because of the complexity, it is difficult to know the relationship among the mean error rate, the sample size, and the feature size. Consequently to analyze the relationship asymptotic expressions are derived and utilized [5, 6]. A procedure to compute the error rate of two-class Gaussian problem with unknown individual covariance matrixes is shown in [7, pp. 91–92]. However the used classifier is not the optimum one and the result is not simple enough to be extended for theoretical analysis of the limited sample effect. Hughes derived and studied the mean error rate over all two-class problems, which does not depend on the unknown population parameters [8]. Since this problem average error rate only depends on the sample size and the measurement complexity, it has been widely studied in theoretical analysis of limited sample effect [9, 10, 11]. However due to the assumption of the uniform prior distributions over the parameter simplices, the problem average error rate can not be applied to evaluate the performance of individual classifiers.

In subsequent sections, a case with unknown covariance matrix (with known mean vector) is first described in Section 2 to 4, and then a case where both parameters are unknown is described in Section 5.

2 Sample conditional density of Gaussian population

Sample conditional density of d-dimensional feature vector X of Gaussian population with unknown covariance matrix given a sample $\chi = \{X_1, X_2, \ldots, X_n\}$ is expressed by

$$p(X|\chi) = \int_S p(X|K)p(K|\chi)dK \ , \tag{1}$$

where K is the inverse of the population covariance matrix and S is $d(d+1)/2$ dimensional subspace on which K is positive definite. Since the mean vector is known, it can be assumed to be zero vector without loss of generality. Then the density $p(X|K)$ is the d-variate Gaussian distribution given by

$$\begin{aligned} p(X|K) &= N(0, K^{-1}) \\ &= (2\pi)^{-\frac{d}{2}}|K|^{\frac{1}{2}}\exp(-\frac{1}{2}X^tKX) \end{aligned} \tag{2}$$

and the density $p(K|\chi)$ is the Wishart distribution of n_n degrees of freedom [2], [7, pp. 392–393] given by

$$\begin{aligned} p(K|\chi) &= W_{n_n}(\Sigma_n) \\ &= c(d, n_n+1)\left|\frac{n_n\Sigma_n}{2}\right|^{\frac{n_n}{2}}|K|^{\frac{n_n-d-1}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(n_n\Sigma_nK)\right\} \end{aligned}$$

$$\begin{aligned} c(d, n_n) &= \left\{\pi^{\frac{d(d-1)}{4}}\prod_{i=1}^{d}\Gamma\left(\frac{n_n-i}{2}\right)\right\}^{-1} \\ \Sigma_n &= \frac{n_0\Sigma_0 + n\Sigma}{n_0+n} \\ \Sigma &= \frac{1}{n}\sum_{i=1}^{n}X_iX_i^t \\ n_n &= n_0 + n \ , \end{aligned} \tag{3}$$

where Σ_0 and n_0 are an initial estimate of the population covariance matrix, and the confidence constant, respectively. When n_0 is set to zero, n_n and Σ_n coincide to n and Σ respectively, and no knowledge about the prior distribution is utilized. $\Gamma(x)$ is the gamma function defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt \ . \tag{4}$$

In [2], [7, pp. 392–393], the density $p(K|\chi)$ is assumed to be the Wishart distribution of $n_n - 1$ degrees of freedom, however it should be replaced by n_n when the population mean vector is known. This point will be examined in Section 6 by computer simulation.

2.1 Univariate case

It is easily shown in univariate case that the conditional density itself is the t-distribution. For univariate case, (1) is written as

$$p(x|\chi) = \int_0^\infty p(x|k)p(k|\chi)dk \ , \tag{5}$$

where k is the reciprocal of the population variance. The density $p(x|k)$ is the univariate Gaussian distribution given by

$$p(x|k) = \frac{1}{\sqrt{2\pi}} k^{\frac{1}{2}} \exp(-\frac{1}{2}kx^2) \tag{6}$$

and the density $p(k|\chi)$ is given by

$$\begin{aligned} p(k|\chi) &= c(1, n_n + 1)\left(\frac{n_n\sigma_n^2}{2}\right)^{\frac{n_n}{2}} k^{\frac{n_n-2}{2}} \exp\left\{-\frac{1}{2}n_n\sigma_n^2 k\right\} \\ &= \frac{1}{\Gamma\left(\frac{n_n}{2}\right)}\left(\frac{n_n\sigma_n^2}{2}\right)^{\frac{n_n}{2}} k^{\frac{n_n-2}{2}} \exp\left\{-\frac{1}{2}n_n\sigma_n^2 k\right\} \end{aligned}$$

$$\sigma_n^2 = \frac{n_0\sigma_0^2 + n\sigma^2}{n_0 + n} \ . \tag{7}$$

By setting $n_n\sigma_n^2 k = r$ the univariate Wishart distribution (7) is transformed to the chi-squared distribution of n_n degrees of freedom $\chi^2_{n_n}(r)$.

$$p(k|\chi)dk = \frac{1}{2^{\frac{n_n}{2}}\Gamma\left(\frac{n_n}{2}\right)} r^{\frac{n_n}{2}-1} e^{-\frac{1}{2}r} dr = \chi^2_{n_n}(r)dr \ . \tag{8}$$

Substituting the (6) and (7) to (5), we have

$$p(x|\chi) = \frac{1}{\sqrt{2\pi}} \frac{(n_n\sigma_n^2)^{\frac{n_n}{2}}}{2^{\frac{n_n}{2}}\Gamma\left(\frac{n_n}{2}\right)} \int_0^\infty k^{\frac{n_n-1}{2}} \exp\left\{-\frac{1}{2}k(x^2 + n_n\sigma_n^2)\right\} dk \ . \tag{9}$$

By setting $k(x^2 + n_n\sigma_n^2)/2 = t$, the integration is performed to lead

$$\begin{aligned} p(x|\chi) &= \frac{1}{\sqrt{2\pi}} \frac{(n_n\sigma_n^2)^{\frac{n_n}{2}}}{2^{\frac{n_n}{2}}\Gamma\left(\frac{n_n}{2}\right)} \left(\frac{2}{x^2 + n_n\sigma_n^2}\right)^{\frac{n_n+1}{2}} \int_0^\infty t^{\frac{n_n-1}{2}} e^{-t} dt \\ &= (n_n\pi)^{-\frac{1}{2}} \sigma_n^{-1} \frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n}{2}\right)} \left\{1 + \frac{1}{n_n}\left(\frac{x}{\sigma_n}\right)^2\right\}^{-\frac{n_n+1}{2}} \end{aligned} \tag{10}$$

Further by setting $x/\sigma_n = y$, we have

$$p(x|\chi)dx = (n_n\pi)^{-\frac{1}{2}} \frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n}{2}\right)} \left(1 + \frac{y^2}{n_n}\right)^{-\frac{n_n+1}{2}} dy \ , \tag{11}$$

where the distribution of y is the t-distribution of nn degrees of freedom $t_{n_n}(y)$. When the population mean is m instead of zero, the conditional density is given by

$$p(x|\chi) = (n_n\pi)^{-\frac{1}{2}}\sigma_n^{-1}\frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n}{2}\right)}\left\{1+\frac{1}{n_n}\left(\frac{x-m}{\sigma_n}\right)^2\right\}^{-\frac{n_n+1}{2}} = t_{n_n}\left(\frac{x-m}{\sigma_n}\right) . \tag{12}$$

The variance of $t_n(y)$ is $n/(n-2)$ for $n > 2$, and the variance of x in (12) is $n_n/(n_n-2)\sigma_n^2$ for $n_n > 2$. Note that the normalizing factor σ_n in (12) is not the square root of the variance of x but is obtained by omitting the factor $n_n/(n_n-2)$. The optimum discriminant function is derived from (12) as

$$\begin{aligned} g(x) &= -2\log\{p(x|\chi)P(\omega)\} \\ &= (n_n+1)\log\left\{1+\frac{1}{n_n}\left(\frac{x-m}{\sigma_n}\right)^2\right\} + \log\sigma_n^2 - 2\log D - 2\log P(\omega) \end{aligned}$$

$$D = (n_n\pi)^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n}{2}\right)} . \tag{13}$$

2.2 Multivariate case

For multivariate case, substituting (2) and (3) to (1), and performing the integration [2], we have

$$p(X|\chi) = (n_n\pi)^{-\frac{d}{2}}|\Sigma_n|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n-d+1}{2}\right)}\left\{1+\frac{1}{n_n}(X-M)^t\Sigma_n^{-1}(X-M)\right\}^{-\frac{n_n+1}{2}} . \tag{14}$$

This result is slightly different from (31) of [2] due to the difference of the degrees of freedom of the Wishart distribution (3).

By setting

$$X - M = \sqrt{\frac{n_n}{n_n-d+1}}T \tag{15}$$

(14) is transformed to

$$p(X|\chi)dX = \{(n_n-d+1)\pi\}^{-\frac{d}{2}}|\Sigma_n|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n-d+1}{2}\right)}\left\{1+\frac{1}{n_n-d+1}T^t\Sigma_n^{-1}T\right\}^{-\frac{n_n+1}{2}}dT , \tag{16}$$

where the distribution of T is the multivariate elliptical t-distribution with n_n-d+1 degrees of freedom [12]. The covariance matrix of T is $(n_n-d+1)/(n_n-d-1)\Sigma_n$ for $n_n > 2$.

3 Optimum discriminant function

1) General case

The optimum discriminant function for general case is derived from (14) as

$$\begin{aligned} g(X) &= -2\log\{p(X|\chi)P(\omega)\} \\ &= (n_n+1)\log\left\{1+\frac{1}{n_n}(X-M)^t\Sigma_n^{-1}(X-M)\right\} + \log|\Sigma_n| - 2\log D - 2\log P(\omega) \end{aligned}$$

$$D = (n_n\pi)^{-\frac{d}{2}}\frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n-d+1}{2}\right)} . \tag{17}$$

2) Case of equal sample size

For a case of equal sample size, the third term D of (17) is common to all classes and is neglected without affection on the decision rule as

$$g(X) = (n_n+1)\log\left\{1+\frac{1}{n_n}(X-M)^t\Sigma_n^{-1}(X-M)\right\} + \log|\Sigma_n| - 2\log P(\omega) . \tag{18}$$

Further transformation leads the following equivalent discriminant function in a quadratic form which yields hyper quadratic decision surface:

$$g(X) = C\left\{1 + \frac{1}{n_n}(X - M)^t \Sigma_n^{-1}(X - M)\right\}$$
$$C = \left\{|\Sigma_n|\, P(\omega)^{-2}\right\}^{\frac{1}{n_n+1}} \tag{19}$$

When sample size of relevant classes are different, the optimum discriminant function is not expressed in quadratic form.

3) Case of equal sample size, equal determinants of sample covariance matrixes, and equal *a priori* probabilities

Because C and n_n is common to the classes, (19) is further simplified to

$$g(X) = (X - M)^t \Sigma_n^{-1}(X - M)\ . \tag{20}$$

4) Case of equal sample size, equal sample covariance matrixes, and equal *a priori* probabilities

(20) is reduced to a linear discriminant function:

$$\begin{aligned} g(X) &= W^t X + w_0 \\ W^t &= -2M^t \Sigma_n^{-1} \\ w_0 &= M^t \Sigma_n^{-1} M \end{aligned} \tag{21}$$

If the sample size or the *a priori* probabilities are not common to the classes the optimum discriminant function is not expressed in linear form even if the sample covariance matrixes are common.

5) Asymptotically optimum discriminant function

As n tends to infinity, Σ_n tends to K^{-1}. Applying Stirling's asymptotic formula to D in (17), we have

$$g(X) = (X - M)^t K(X - M) + \log\left|K^{-1}\right| - 2\log P(\omega)\ , \tag{22}$$

which is the optimum discriminant function (quadratic discriminant function) for Gaussian distributions with known parameters.

4 Evaluation of mean error rate

4.1 Univariate case

4.1.1 Case with common sample size

Since the discriminant function for a case of equal sample size is univariate quadratic, the decision boundary and the mean error rate are easily calculated. For simplicities sake, the *a priori* probabilities are assumed to be common to classes. The extension to unequal *a priori* probability case is straight forward.

The discriminant function

$$g_i(x) = \sigma_{ni}^{\frac{2}{n_n+1}}\left\{1 + \frac{1}{n_n}\left(\frac{x - m_i}{\sigma_{ni}}\right)^2\right\} \qquad (i = 1, 2) \tag{23}$$

is derived from (13) or (19). Setting $h(x) = 0$ for

$$\begin{aligned} h(x) &= g_1(x) - g_2(x) \\ &= \sigma_{n1}^{\frac{2}{n_n+1}}\left\{1 + \frac{1}{n_n}\left(\frac{x - m_1}{\sigma_{n1}}\right)^2\right\} - \sigma_{n2}^{\frac{2}{n_n+1}}\left\{1 + \frac{1}{n_n}\left(\frac{x - m_2}{\sigma_{n2}}\right)^2\right\} \\ &= (a - b)x^2 - 2(am_1 - bm_2)x + am_1^2 - bm_2^2 + c \end{aligned}$$

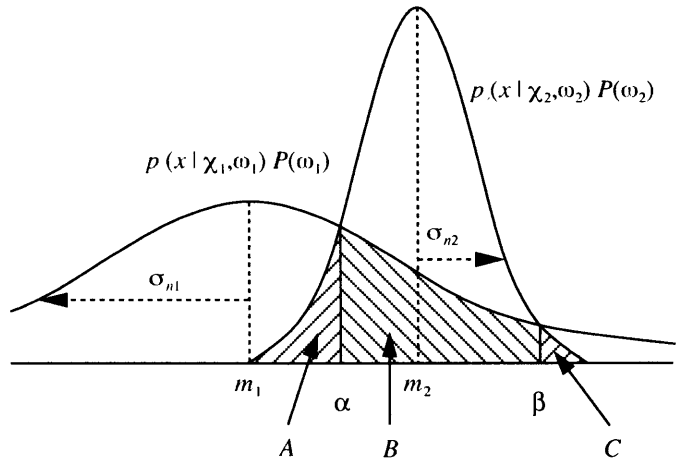

Figure 1: Calculation of mean error rate (univariate case).

$$a = \frac{1}{n_n}\sigma_{n1}^{\frac{2}{n_n+1}-2}, \qquad b = \frac{1}{n_n}\sigma_{n2}^{\frac{2}{n_n+1}-2}, \qquad c = \sigma_{n1}^{\frac{2}{n_n+1}} - \sigma_{n2}^{\frac{2}{n_n+1}} \tag{24}$$

The decision boundaries are determined as

$$\begin{aligned} \alpha &= \beta = \frac{m_1+m_2}{2} && (\sigma_{n1} = \sigma_{n2}) \\ \alpha, \beta &= \frac{am_1 - bm_2 \mp \sqrt{(m_1-m_2)^2 ab - (a-b)c}}{a-b} && (\sigma_{n1} \neq \sigma_{n2}) \ . \end{aligned} \tag{25}$$

The mean error rate for $\sigma_{n1} \geq \sigma_{n2}$ is given by

$$\begin{aligned} \varepsilon &= P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2 = P(\omega_1)P(error|\chi,\omega_1) + P(\omega_2)P(error|\chi,\omega_2) \\ &= B + A + C \end{aligned}$$

$$\begin{aligned} A &= \int_{-\infty}^{\alpha} p(x|\chi,\omega_2)P(\omega_2)dx = \frac{1}{2}\Phi_{n_n}\left(\frac{\alpha - m_2}{\sigma_{n2}}\right) \\ B &= \int_{\alpha}^{\beta} p(x|\chi,\omega_1)P(\omega_1)dx = \frac{1}{2}\Phi_{n_n}\left(\frac{\beta - m_1}{\sigma_{n1}}\right) - \frac{1}{2}\Phi_{n_n}\left(\frac{\alpha - m_1}{\sigma_{n1}}\right) \\ C &= \int_{\beta}^{\infty} p(x|\chi,\omega_2)P(\omega_2)dx = \frac{1}{2}\left\{1 - \Phi_{n_n}\left(\frac{\beta - m_2}{\sigma_{n2}}\right)\right\} \end{aligned} \tag{26}$$

(Figure 1), where $\Phi_n(x_0)$ is defined by

$$\Phi_n(x_0) = \int_{-\infty}^{x_0} t_n(x)dx \ . \tag{27}$$

Using the incomplete beta function

$$I_{x_0}(p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \int_0^{x_0} x^{p-1}(1-x)^{q-1}dx \tag{28}$$

$\Phi_n(x_0)$ is given by

$$\Phi_n(x_0) = 1 - \frac{1}{2} I_{\frac{n}{n+x_0^2}}\left(\frac{n}{2}, \frac{1}{2}\right) \ . \tag{29}$$

The mean error rate obtained by the quadrature of the conditional density has an interpretation shown by

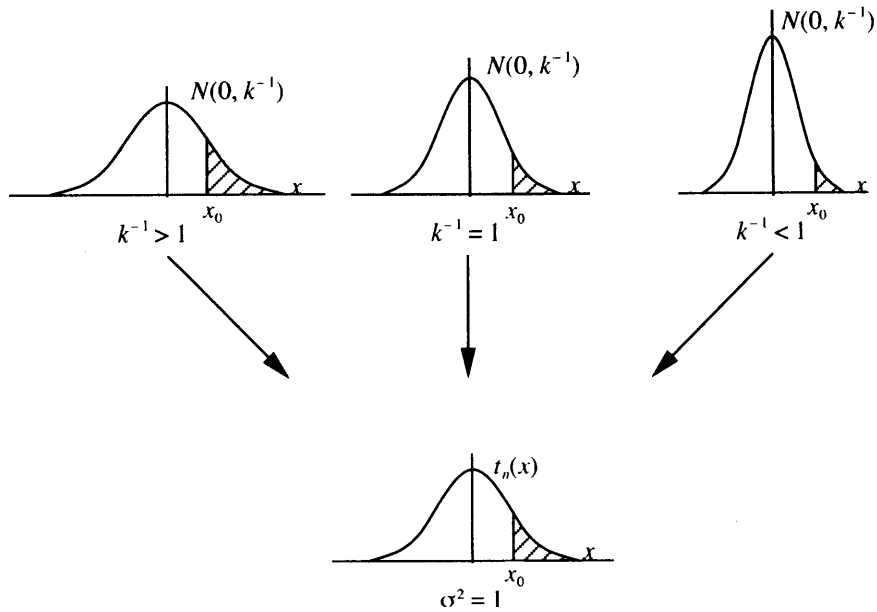

Figure 2: Relationship between the error rate for individual population with variance k^{-1} and the expectation.

$$\begin{aligned}\int_{x_0}^{\infty} p(x|\chi)dx &= \int_{x_0}^{\infty}\int_{0}^{\infty} p(x|k)p(k|\chi)dkdx \\ &= \int_{0}^{\infty}\int_{x_0}^{\infty} p(x|k)dxp(k|\chi)dk = \int_{0}^{\infty} P(error|k)p(k|\chi)dk \ , \end{aligned} \tag{30}$$

where $p(x|\chi)$ is a sample conditional density of a class, x_0 is a decision boundary, and the region of the class is assumed to be $[-\infty, x_0)$. This equation implies that the mean error rate is the expectation of the error rate $p(error|k)$ given unknown population parameter k. Figure 2 shows the relationship between the error rate for individual population with variance k^{-1} and the expectation, when the sample variance is 1. When sample size is n and $n_0 = 0$, nk is subject to the chi-squared distribution with n degrees of freedom as shown by (8), and the sample conditional density is the t-distribution with n degrees of freedom. The expectation of the hatched area of Gaussian distributions in upper row is equal to the hatched area of the t-distribution in lower row. This relation also holds for cases where there are multiple decision boundaries and the region of a class is separated into multiple segments.

4.1.2 Case with different sample size

When the sample size of relevant classes are different, the optimum discriminant function is given by

$$\begin{aligned} g_i(x) &= \left(\frac{\sigma_{ni}}{D_i}\right)^2 \left\{1 + \frac{1}{n_{ni}}\left(\frac{x - m_i}{\sigma_{ni}}\right)^2\right\}^{n_{ni}+1} \\ D_i &= (n_{ni}\pi)^{-\frac{1}{2}} \frac{\Gamma\left(\frac{n_{ni}+1}{2}\right)}{\Gamma\left(\frac{n_{ni}}{2}\right)} \qquad (i = 1, 2) \ . \end{aligned} \tag{31}$$

The *a priori* probabilities are assumed to be common to all classes. The extension to unequal *a priori* probability case is straight forward.

Unfortunately the equation $h(x) = 0$ can not be solved analitically. In the experiment in Section 6, the decision boundary is calculated employing Newton's iterative formula:

$$x_{k+1} = x_k - \frac{h(x_k)}{h'(x_k)}$$

$$
\begin{aligned}
h(x) &= \left(\frac{\sigma_{n1}}{D_1}\right)^2 \left\{1+\frac{1}{n_{n1}}\left(\frac{x-m_1}{\sigma_{n1}}\right)^2\right\}^{n_{n1}+1} - \left(\frac{\sigma_{n2}}{D_2}\right)^2 \left\{1+\frac{1}{n_{n2}}\left(\frac{x-m_2}{\sigma_{n2}}\right)^2\right\}^{n_{n2}+1} \\
h'(x) &= \frac{2(n_{n1}+1)}{n_{n1}\sigma_{n1}}\left(\frac{x-m_1}{\sigma_{n1}}\right)\left(\frac{\sigma_{n1}}{D_1}\right)^2 \left\{1+\frac{1}{n_{n1}}\left(\frac{x-m_1}{\sigma_{n1}}\right)^2\right\}^{n_{n1}} \\
&\quad - \frac{2(n_{n2}+1)}{n_{n2}\sigma_{n2}}\left(\frac{x-m_2}{\sigma_{n2}}\right)\left(\frac{\sigma_{n2}}{D_2}\right)^2 \left\{1+\frac{1}{n_{n2}}\left(\frac{x-m_{n2}}{\sigma_{n2}}\right)^2\right\}^{n_{n2}}
\end{aligned} \tag{32}
$$

The mean error rate is calculated in the similar way as (26).

4.2 Multivariate case

4.2.1 Conditional error rate given population covariance matrixes

The error rate for specified population is evaluated using the population covariance matrix. The sample size, the covariance matrixes and the *a priori* probabilities are assumed to be common to two classes. The logarithm of the likelihood ratio is given by

$$
\begin{aligned}
h(X) &= \frac{1}{2}(X-M_1)^t\Sigma_n^{-1}(X-M_1) - \frac{1}{2}(X-M_2)^t\Sigma_n^{-1}(X-M_2) \\
&= (M_2-M_1)^t\Sigma_n^{-1}X + \frac{1}{2}(M_1^t\Sigma_n^{-1}M_1 - M_2^t\Sigma_n^{-1}M_2)\ .
\end{aligned} \tag{33}
$$

When the population covariance matrix K^{-1} is given, the distribution of $X$is the multivariate Gaussian, and the distribution of the linear mapping $h(X)$ is the univariate Gaussian. The means of $h(X)$ are given by

$$
\eta_i = (M_2-M_1)^t\Sigma_n^{-1}E\{X|\omega_i\} + \frac{1}{2}(M_1^t\Sigma_n^{-1}M_1 - M_2^t\Sigma_n^{-1}M_2) \tag{34}
$$

$(i=1,2)$, which leads to

$$
\begin{aligned}
\eta_1 &= -\frac{1}{2}(M_2-M_1)^t\Sigma_n^{-1}(M_2-M_1) \\
\eta_2 &= -\eta_1 = \eta\ .
\end{aligned} \tag{35}
$$

The variances of $h(X)$ are given by

$$
\sigma_i^2 = E[\{h(X)-\eta_i\}^2|\omega_i] \tag{36}
$$

$(i=1,2)$, which leads to

$$
\begin{aligned}
\sigma_1^2 &= (M_2-M_1)^t\Sigma_n^{-1}E\{(X-M_1)(X-M_1)^t|\omega_1\}\Sigma_n^{-1}(M_2-M_1) \\
&= (M_2-M_1)^t\Sigma_n^{-1}K_1^{-1}\Sigma_n^{-1}(M_2-M_1) \\
\sigma_2^2 &= (M_2-M_1)^t\Sigma_n^{-1}K_2^{-1}\Sigma_n^{-1}(M_2-M_1)\ .
\end{aligned} \tag{37}
$$

Using these parameters, the conditional error rate is given by

$$
\varepsilon = P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2 = \frac{1}{2}\Phi\left(\frac{\eta_1}{\sigma_1}\right) + \frac{1}{2}\left\{1-\Phi\left(\frac{\eta_2}{\sigma_2}\right)\right\} \tag{38}
$$

(Figure 3), where $\Phi(x_0)$ is the Gaussian error function

$$
\Phi(x_0) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x_0}\exp(-\frac{1}{2}x^2)dx\ . \tag{39}
$$

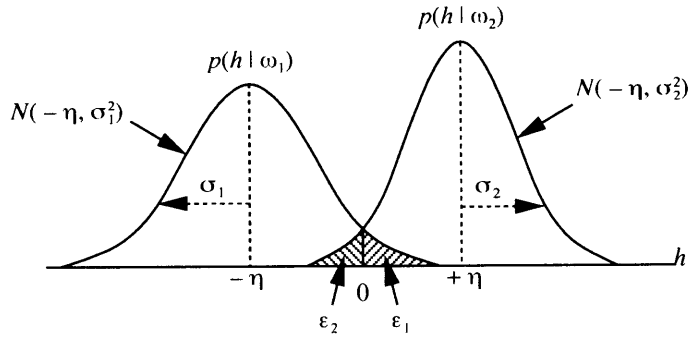

Figure 3: Density functions of $h(X)$ and mean error rate of two-class problem with common sample covariance matrixes and known individual population covariance matrixes.

4.2.2 Mean error rate over population covariance matrixes

When the population covariance matrixes are random parameters, we use next $h(X)$ instead of (33).

$$h(X) = (M_2 - M_1)^t \Sigma_n^{-1} \sqrt{\frac{n_n - d + 1}{n_n}} X + \frac{1}{2}\sqrt{\frac{n_n - d + 1}{n_n}} (M_1^t \Sigma_n^{-1} M_1 - M_2^t \Sigma_n^{-1} M_2) \tag{40}$$

The distribution of $((n_n - d + 1)/n_n)^{1/2} X$ is d-variate elliptical t-distribution with $n_n - d + 1$ degrees of freedom, and the distribution of $h(X)$ is univariate t-distribution with the same degrees of freedom. Similar to 4.2.1, the means of $h(X)$ are given by

$$\begin{aligned} \eta_1 &= -\frac{1}{2}\sqrt{\frac{n_n - d + 1}{n_n}} \delta_n^2 \\ \eta_2 &= \frac{1}{2}\sqrt{\frac{n_n - d + 1}{n_n}} \delta_n^2 \\ & \delta_n^2 = (M_2 - M_1)^t \Sigma_n^{-1} (M_2 - M_1) \ . \end{aligned} \tag{41}$$

The variances of $h(X)$ is given by

$$\begin{aligned} \sigma_1^2 &= (M_2 - M_1)^t \Sigma_n^{-1} E\{\frac{n_n - d + 1}{n_n}(X - M_1)(X - M_1)^t | \omega_1\} \Sigma_n^{-1} (M_2 - M_1) \\ &= \frac{n_n - d + 1}{n_n - d - 1}(M_2 - M_1)^t \Sigma_n^{-1} (M_2 - M_1) = \frac{n_n - d + 1}{n_n - d - 1}\delta_n^2 \\ \sigma_2^2 &= \frac{n_n - d + 1}{n_n - d - 1}\delta_n^2 \ . \end{aligned} \tag{42}$$

Using these parameters the mean error rate is given by

$$\begin{aligned} \varepsilon &= P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2 \\ &= \frac{1}{2}\Phi_{n_n - d + 1}\left(\frac{\eta_1}{\delta_n}\right) + \frac{1}{2}\left\{1 - \Phi_{n_n - d + 1}\left(\frac{\eta_2}{\delta_n}\right)\right\} \\ &= 1 - \Phi_{n_n - d + 1}\left(\frac{1}{2}\sqrt{\left(1 - \frac{d-1}{n_n}\right)\delta_n^2}\right) \ . \end{aligned} \tag{43}$$

When $n_0 = 0$,

$$\varepsilon = 1 - \Phi_{n - d + 1}\left(\frac{1}{2}\sqrt{\left(1 - \frac{d-1}{n}\right)\delta^2}\right) \ . \tag{44}$$

The formulas of the mean error rate (43) and (44) have the following characteristics when compared with the formulas ever known.

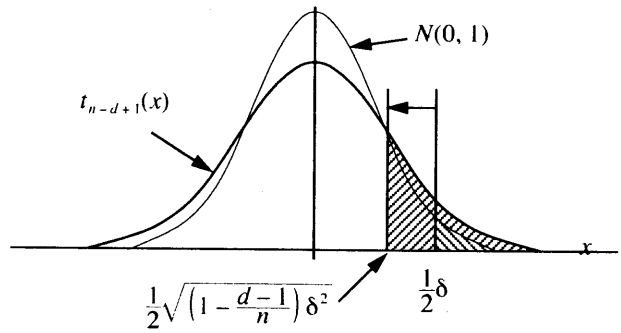

Figure 4: Increase of mean error rate due to limited sample effect.

1) The unknown population parameters are not required in its calculation.
2) The expression is simple and clearly shows the limited sample effect on the mean error rate.
3) The relationship between the prior parameters n_0, Σ_0 and the mean error rate is explicitly expressed.

It should be noted that the Mahalanobis distance δ in (44) is an apparent one which is calculated using the known population mean vector and the sample covariance matrix. (44) reveals two causes which increase the mean error rate due to the limited sample effect. One is that the area of the tail of t-distribution increases due to the reduction of the degrees of freedom. The other is that the apparent squared Mahalanobis distance between two classes shrinks by $(d-1)/n$, and increases the mean error rate (Figure 4). The affection of the former is marginal and is negligible if $n-d+1$ is greater than 20 or so, because the t-distribution with this degrees of freedom can be approximated by the Gaussian distribution, which is the t-distribution with infinite degrees of freedom. On the other hand, the affection of the latter is so severe and is not negligible unless the sample size is much larger than the dimensionality. Such shrinkage of the apparent Mahalanobis distance has its origin in the variable transformation by (15), and causes a mysterious problem so called "peaking phenomenon" or "curse of dimensionality". This undesirable phenomenon is caused and aggravated by neglecting the prior distribution by setting $n_0 = 0$. The case for $n_0 \neq 0$ is discussed in Section 7.

5 Case with unknown mean vectors and unknown covariance matrixes

5.1 Conditional density and optimum classifier

When the mean vectors and the covariance matrixes are both unknown, the conditional density of X is given by

$$p(X|\chi) = \int_S \int p(X|M,K)p(M,K|\chi)dMdK \ , \tag{45}$$

where $p(X|M,K)$ is d-variate Gaussian distribution given by

$$\begin{aligned} p(X|M,K) &= N(M,K^{-1}) \\ &= (2\pi)^{-\frac{d}{2}}|K|^{\frac{1}{2}}\exp\{-\frac{1}{2}(X-M)^tK(X-M)\} \end{aligned} \tag{46}$$

and $p(M,K|\chi)$ is the Gauss-Wishart distribution with $n_n - 1$ degrees of freedom [2], [7, pp. 392–393] given by

$$\begin{aligned} p(M,K|\chi) &= W_{n_n-1}(\Sigma_n) \\ &= (2\pi)^{-\frac{d}{2}}|w_0K|^{\frac{1}{2}}\exp\{-\frac{1}{2}w_0(M-\mu_n)^tK(M-\mu_n)\} \\ &\quad \cdot c(d,n_n)\left|\frac{n_n\Sigma_n}{2}\right|^{\frac{n_n-1}{2}}|K|^{\frac{n_n-d-2}{2}}\exp\left\{-\frac{1}{2}\mathrm{tr}(n_n\Sigma_nK)\right\} \end{aligned}$$

$$
\begin{aligned}
\mu_n &= \frac{w_0\mu_0 + n\mu}{w_0 + n} \\
\Sigma_n &= \frac{(n_0\Sigma_0 + w_0\mu_0\mu_0^t) + \{(n-1)\sum + n\mu\mu^t\} - w_n\mu_n\mu_n^t}{n_0 + n} \\
n_n &= n_0 + n \\
w_n &= w_0 + n \\
\mu &= \frac{1}{n}\sum_{i=1}^{n} X_i \\
\Sigma &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)(X_i - \mu)^t \ ,
\end{aligned} \tag{47}
$$

where μ_0 and $w0$ are an initial estimate of the population mean vector and its confidence constant, respectively. Substituting (46) and (47) to (45) and performing the integration [2], we have

$$
p(X|\chi) = (n_n\pi)^{-\frac{d}{2}}|\Sigma_n|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n_n}{2}\right)}{\Gamma\left(\frac{n_n-d}{2}\right)}\left(\frac{w_n+1}{w_n}\right)^{\frac{d}{2}}\left\{1 + \frac{w_n+1}{w_n}\frac{1}{n_n}(X-\mu_n)^t\Sigma_n^{-1}(X-\mu_n)\right\}^{-\frac{n_n}{2}} . \tag{48}
$$

The term $(w_n+1)/{w_n}^{d/2}$ in (48) is missing in (50) of [2].
By setting

$$
X - \mu_n = \sqrt{\frac{w_n}{w_n+1}\frac{n_n}{n_n-d}}T \tag{49}
$$

(48) leads to

$$
p(X|\chi)dX = \{(n_n-d)\pi\}^{-\frac{d}{2}}|\Sigma_n|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n_n}{2}\right)}{\Gamma\left(\frac{n_n-d}{2}\right)}\left\{1 + \frac{1}{n_n-d}T^t{\Sigma_n}^{-1}T\right\}^{-\frac{n_n}{2}}dT . \tag{50}
$$

The distribution of T is d-variate elliptical t-distribution with $n_n - d$ degrees of freedom. From (48), the optimum discriminant function is derived as

$$
\begin{aligned}
g(X) &= -2\log\{p(X|\chi)P(\omega)\} \\
&= n_n\log\left\{1 + \frac{w_n+1}{w_n}\frac{1}{n_n}(X-\mu_n)^t\Sigma_n^{-1}(X-\mu_n)\right\} + \log|\Sigma_n| - 2\log D - 2\log P(\omega) \\
D &= \left(\frac{w_n}{w_n+1}n_n\pi\right)^{-\frac{d}{2}}\frac{\Gamma\left(\frac{n_n}{2}\right)}{\Gamma\left(\frac{n_n-d}{2}\right)} .
\end{aligned} \tag{51}
$$

5.2 Evaluation of mean error rate

The mean error rate is obtained as the expectation over the random population parameters. The sample size, the sample covariance matrixes, and the *a priori* probabilities are assumed to be common to two classes.

We use next $h(X)$ instead of (40).

$$
h(X) = (\mu_{n2} - \mu_{n1})^t\Sigma_n^{-1}\sqrt{\frac{(n_n-d)(w_n+1)}{n_nw_n}}X + \frac{1}{2}\sqrt{\frac{(n_n-d)(w_n+1)}{n_nw_n}}(\mu_{n1}^t\Sigma_n^{-1}\mu_{n1} - \mu_{n2}^t\Sigma_n^{-1}\mu_{n2}) \tag{52}
$$

Similar to 4.2.2, the distribution of $h(X)$ is univariate t-distribution with $n_n - d$ degrees of freedom, and the means $\eta_i (i = 1, 2)$ are given by

$$
\begin{aligned}
\eta_1 &= -\frac{1}{2}\sqrt{\frac{(n_n-d)(w_n+1)}{n_nw_n}}\delta_n^2 \\
\eta_2 &= \frac{1}{2}\sqrt{\frac{(n_n-d)(w_n+1)}{n_nw_n}}\delta_n^2 \\
\delta_n^2 &= (\mu_{n2} - \mu_{n1})^t\Sigma_n^{-1}(\mu_{n2} - \mu_{n1}) .
\end{aligned} \tag{53}
$$

The variances $\sigma_i^2 (i = 1, 2)$ of $h(X)$ are given by

$$\begin{aligned}
\sigma_1^2 &= (\mu_{n2} - \mu_{n1})^t \Sigma_n^{-1} E\left\{ \frac{(n_n - d)(w_n + 1)}{n_n w_n} (X - \mu_{n1})(X - \mu_{n1})^t | \omega_1 \right\} \Sigma_n^{-1} (\mu_{n2} - \mu_{n1}) \\
&= \frac{n_n - d}{n_n - d - 2} \delta_n^2 \\
\sigma_2^2 &= \frac{n_n - d}{n_n - d - 2} \delta_n^2 \ . \qquad (54)
\end{aligned}$$

Using these parameters the mean error rate ε is given by

$$\varepsilon = 1 - \Phi_{n_n - d}\left(\frac{1}{2} \sqrt{\frac{w_n + 1}{w_n} \left(1 - \frac{d}{n_n}\right) \delta_n^2} \right) \ . \qquad (55)$$

When $n_0 = w_0 = 0$, it is given by

$$\varepsilon = 1 - \Phi_{n-d}\left(\frac{1}{2} \sqrt{\frac{n + 1}{n} \left(1 - \frac{d}{n}\right) \delta^2} \right) \ . \qquad (56)$$

6 Computer simulation

6.1 Bayesian sampling

In the following computer simulation, a new sampling procedure called Bayesian sampling is employed together with the ordinary sampling procedure. Figure 5 illustrates the relationship between the ordinary sampling (a) and the Bayesian sampling (b). In the ordinary sampling, specified size of sample are drawn from a specified population and the sample parameters are calculated. Figure 5 (a) illustrates the case with a Gaussian population $N(0, I)$ and three samples of size five with the sample covariance matrixes Σ_a, Σ_b, and Σ_c. The classifiers are designed using these sample parameters and the mean error rate for the population is evaluated. Since the sample parameters are random variables, the expectation of the error rate is taken by repeating the sampling and the design and test of the classifier. On the contrary the Bayesian sampling generates populations from which a sample with specified parameter, e.g. $N(0, I)$, is extracted. When a sample of specified size is drawn from a temporal population $N(0, I)$, and the sample covariance matrix is Σ_a, the actual population is determined to be $N(0, \Sigma_a^{-1})$. Since the population parameters are random variables in this case, the expectation of the error rate is taken by repeating the Bayesian sampling and the test of the classifier. The design of the classifier need not be repeated because the design sample is fixed through the experiment. In this example, the sample mean vector and the sample covariance matrix are assumed to be zero vector and identity matrix, respectively. The general procedure is described below.

The population parameters are determined so that the parameters of a sample drawn from the population is (μ_2, Σ_2). The parameters of a sample of size n drawn from a temporal population $N(0, I)$ are denoted by (μ_1, Σ_1), i.e.

$$\begin{aligned}
\mu_1 &= \frac{1}{n} \sum_{i=1}^{n} X_i \\
\Sigma_1 &= \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu_1)(X_i - \mu_1)^t \ . \qquad (57)
\end{aligned}$$

By setting

$$Y = \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t (X - \mu_1) \qquad (\Sigma_1 \Phi_1 = \Phi_1 \Lambda_1) \qquad (58)$$

the sample parameters are transformed to $(0, I)$, i.e.

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} Y_i = 0 \\
&\frac{1}{n-1} \sum_{i=1}^{n} Y_i Y_i^t = \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \Sigma_1 \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t = \Phi_1 \Phi_1^t = I \qquad (59)
\end{aligned}$$

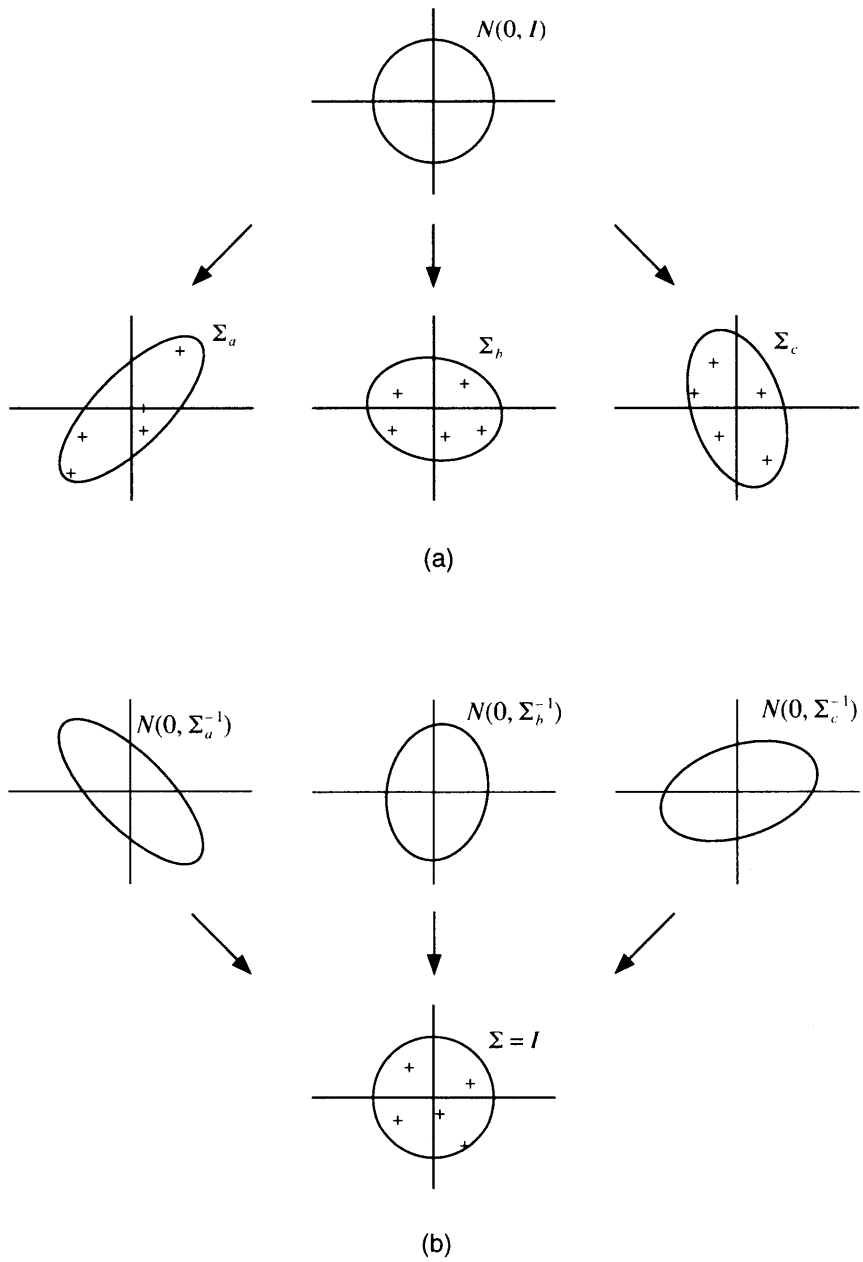

Figure 5: Relationship between ordinary sampling (a), and Bayesian sampling (b).

and the population parameters of Y are given by

$$\begin{aligned} E(Y) &= -\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t\mu_1 \\ V(Y) &= E\left[\{Y-E(Y)\}\{Y-E(Y)\}^t\right] = \Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t = \Sigma_1^{-1}\ , \end{aligned} \tag{60}$$

where Λ_1 and Φ_1 are the eigenvalue matrix and eigenvector matrix of Σ_1, respectively.

Further by setting

$$\begin{aligned} Z &= \Phi_2\Lambda_2^{\frac{1}{2}}Y + \mu_2 \\ &= \Phi_2\Lambda_2^{\frac{1}{2}}\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^tX + \mu_2 - \Phi_2\Lambda_2^{\frac{1}{2}}\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t\mu_1 \quad (\Sigma_2\Phi_2 = \Phi_2\Lambda_2) \end{aligned} \tag{61}$$

the sample parameters are transformed to (μ_2, Σ_2), i.e.

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} Z_i = \Phi_2\Lambda_2^{\frac{1}{2}}\frac{1}{n}\sum_{i=1}^{n} Y_i + \mu_2 = \mu_2 \\ &\frac{1}{n-1}\sum_{i=1}^{n} Y_iY_i{}^t = \Phi_2\Lambda_2^{\frac{1}{2}}I\Lambda_2^{\frac{1}{2}}\Phi_2^t = \Phi_2\Lambda_2\Phi_2^t = \Sigma_2 \end{aligned} \tag{62}$$

and the population parameters of Z are given by

$$\begin{aligned} E(Z) &= \mu_2 - \Phi_2\Lambda_2^{\frac{1}{2}}\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t\mu_1 \\ V(Z) &= \Phi_2\Lambda_2^{\frac{1}{2}}\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t\Phi_1\Lambda_1^{-\frac{1}{2}}\Phi_1^t\Lambda_2^{\frac{1}{2}}\Phi_2^t = \Phi_2\Lambda_2^{\frac{1}{2}}\Sigma_1^{-1}\Lambda_2^{\frac{1}{2}}\Phi_2^t\ . \end{aligned} \tag{63}$$

When the population mean vector M is known, (63) is replaced by

$$\begin{aligned} E(Z) &= M \\ V(Z) &= \Phi_2\Lambda_2^{\frac{1}{2}}\Sigma_1^{-1}\Lambda_2^{\frac{1}{2}}\Phi_2^t \\ &\quad \Sigma_1 = \frac{1}{n}\sum_{i=1}^{n} X_iX_i{}^t \end{aligned} \tag{64}$$

In the following experiments, n_0 is set to zero and the population is assumed to have known mean vector and unknown covariance matrix.

6.1.1 Univariate case

Table 1 to 3 show the experimental results for univariate case. Table 1 shows the mean error rate for a case of equal variances ($\sigma_1^2 = \sigma_2^2 = 1.0$), and Table 2 for different variances ($\sigma_1^2 = 4.0, \sigma_2^2 = 0.25$). The *a priori* probabilities are common to two classes, and the two means are -1.0 and 1.0, respectively. In the tables n_1 and n_2 denote the sample size of each class, and the columns *opt.* and *qdf.* show the mean error rate of the optimum classifier and the quadratic classifier, respectively. The rows *sim.* show the result by Monte Carlo simulation, the rows *Gau.* by Gaussian quadrature using population variance, and the rows t by quadrature of t-distribution. In the Monte Carlo simulation, the Bayesian sampling was employed and 2000 tests each of which used a test sample of size 1000 were repeated to calculate the mean error rate. The error rate by Gaussian quadrature was calculated for each test using the population variance and averaged over 2000 tests. When the sample variance and the sample size are common to classes, the optimum classifier and the quadratic classifier give the same results. For the rest of the case, the optimum classifier always outperforms the quadratic classifier. The mean error rates calculated by three different ways including the theoretical prediction by t-quadrature are accurately coincident mutually. Figure 6 shows the relationship between the mean error rate and the sample size of class1 when total sample size is fixed to 4 in Table 1. When the sample variances are common to two classes, the mean error rate of the quadratic classifier is minimized for the case of equal sample size, while the one of the optimum classifier is maximized for the case. The similar relationship for Table 2 is shown in Figure 7. When the sample variances are not common, the mean error rates of the both classifiers are lower for the case

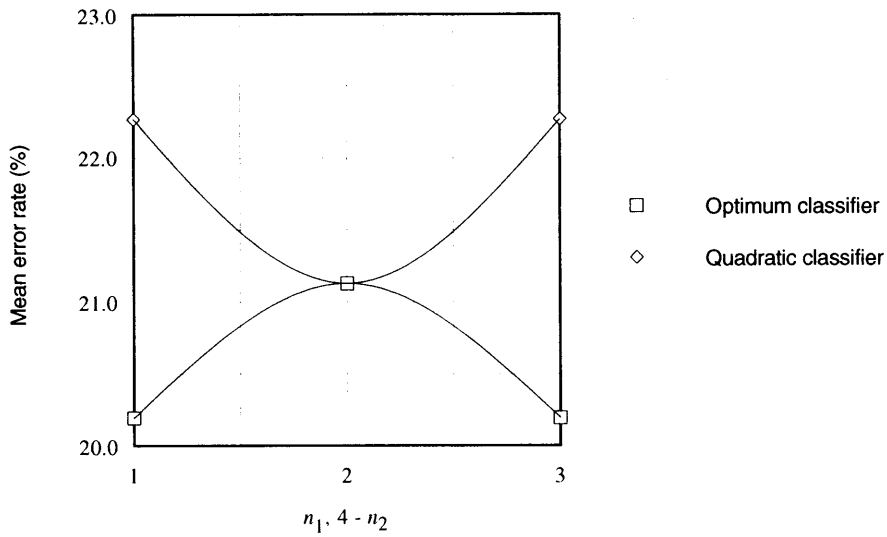

Figure 6: Theoretical mean error rate (%) v.s. sample size with fixed total sample size ($n_1 + n_2 = 4$, $\sigma_1^2 = \sigma_2^2 = 1.0$).

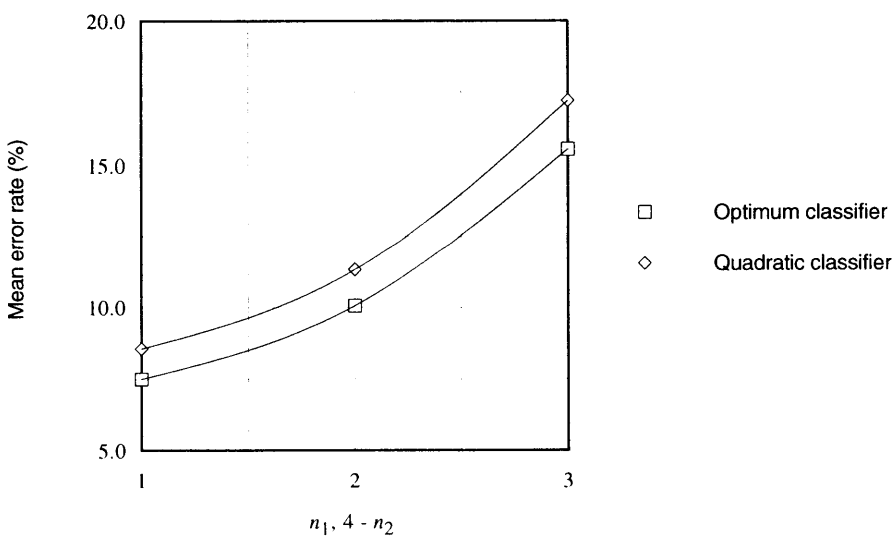

Figure 7: Theoretical mean error rate (%) v.s. sample size with fixed total sample size ($n_1 + n_2 = 4$, $\sigma_1^2 = 4.0$, $\sigma_2^2 = 0.25$).

Table 1: Mean error rate (%) v.s. sample size in univariate two-class problem with common sample variances ($\sigma_1^2 = \sigma_2^2 = 1.0$).

n_1 \ n_2		1		2		3	
		opt.	*qdf.*	*opt.*	*qdf.*	*opt.*	*qdf.*
1	*sim.*	24.98	24.98	21.62	23.06	20.14	22.26
	Gau.	24.99	24.99	21.63	23.07	20.15	22.27
	t	25.00	25.00	21.67	23.07	20.19	22.27
2	*sim.*	21.77	23.08	21.16	21.16	20.13	20.37
	Gau.	21.76	23.08	21.16	21.16	20.13	20.37
	t	21.67	23.07	21.13	21.13	20.17	20.34
3	*sim.*	20.21	22.27	20.19	20.35	19.56	19.56
	Gau.	20.21	22.27	20.19	20.35	19.56	19.56
	t	20.19	22.27	20.17	20.34	19.55	19.55

Table 2: Mean error rate (%) v.s. sample size in univariate two-class problem with individual sample variances ($\sigma_1^2 = 4.0, \sigma_2^2 = 0.25$).

n_1 \ n_2		1		2		3	
		opt.	*qdf.*	*opt.*	*qdf.*	*opt.*	*qdf.*
1	*sim.*	13.97	16.32	9.08	10.68	7.46	8.56
	Gau.	13.99	16.32	9.10	10.68	7.48	8.56
	t	14.00	16.34	9.13	10.70	7.50	8.57
2	*sim.*	15.06	16.93	10.01	11.28	8.31	9.16
	Gau.	15.07	16.93	10.03	11.29	8.32	9.17
	t	15.10	16.96	10.07	11.32	8.35	9.19
3	*sim.*	15.39	17.20	10.39	11.55	8.66	9.43
	Gau.	15.51	17.19	10.40	11.55	8.67	9.43
	t	15.53	17.21	10.43	11.57	8.69	9.45

where the class with less sample variance has greater sample size.In contrast with Table 1 and 2 obtained by Bayesian sampling, Table 3 shows the result of experiment employing the ordinary sampling (non-Bayesian sampling) with different population variances ($k_1^{-1} = 4.0$, $k_2^{-1} = 0.25$). Since the mean error rate obtained by non-Bayesian sampling is not calculated by the t-quadrature, the result is different from those by Monte Carlo simulation and the Gaussian quadrature. However, the optimum classifier always outperforms the quadratic classifier. This result indicates that the optimum classifier is independently optimum from the sampling procedures.

6.1.2 Multivariate case with common sample covariance matrix

Table 4 and 5 show the results of experiments for multivariate case where the sample size, the sample covariance matrixes, and the *a priori* probabilities are all common to two classes. The rows *qdf.* and *opt.* are the results by the Monte Carlo simulation employing the Bayesian sampling, where the size of test sample is 1000, and the number of iteration is 5000. The rows *Gau.* and t shows the mean error rate calculated as described in **4.2.1** and **4.2.2**, respectively. The optimum discriminant function employed in the simulation is (18), which is equivalent to (21) in this case. Table 4 shows the mean error rate when the sample size n increases proportional to the dimensionality $d-1$. Under this condition, the shrinkage of the apparent squared Mahalanobis distance $(d-1)/n$ is fixed, while the degrees of freedom $n-d+1$ of the t-distribution increases. Since the tail of t-distribution is shortened with the increase of the degrees of freedom, the mean error rate reduces with the increase of the dimensionality. The sample covariance matrix is $d \times d$ identity matrix, and the population mean vectors are

$$\begin{aligned} M_1 &= (-1, 0, \cdots, 0), \\ M_2 &= -M_1 \end{aligned} \tag{65}$$

Table 3: Mean error rate (%) v.s. sample size in univariate two-class problem with individual population variances ($k_1^{-1} = 4.0, k_2^{-1} = 0.25$) for non-Bayesian sampling.

n_1 \ n_2		1		2		3	
		opt.	*qdf.*	*opt.*	*qdf.*	*opt.*	*qdf.*
1	*sim.*	12.09	16.51	9.45	11.92	8.43	10.45
	Gau.	12.09	16.50	9.43	11.89	8.40	10.42
	t	10.64	13.61	7.58	9.14	6.43	7.49
2	*sim.*	10.89	15.28	8.81	10.46	8.01	8.92
	Gau.	10.86	15.27	8.78	10.43	7.98	8.88
	t	12.12	14.99	8.89	10.30	7.66	8.52
3	*sim.*	10.44	15.22	8.63	10.51	7.91	8.94
	Gau.	10.42	15.20	8.61	10.47	7.89	8.90
	t	12.68	15.43	9.37	10.71	8.12	8.94

Table 4: Mean error rate (%) v.s. dimensionality in multivariate two-class problem with common sample covariance matrixes and increasing sample size proportional to dimensionality -1.

d \ n		$n = 2(d-1)$		$n = 3(d-1)$		$n = 4(d-1)$	
2	*qdf.*	30.48	$n-d+1=1$	25.05	$n-d+1=2$	22.51	$n-d+1=3$
	opt.	30.48	$(d-1)/n=1/2$	25.05	$(d-1)/n=1/3$	22.51	$(d-1)/n=1/4$
	Gau.	30.46		25.03		22.50	
	t	30.41		25.00		22.51	
3	*qdf.*	27.61	$n-d+1=2$	22.90	$n-d+1=4$	20.92	$n-d+1=6$
	opt.	27.61	$(d-1)/n=1/2$	22.90	$(d-1)/n=1/3$	20.92	$(d-1)/n=1/4$
	Gau.	27.62		22.91		20.92	
	t	27.64		23.00		20.99	
4	*qdf.*	26.60	$n-d+1=3$	22.30	$n-d+1=6$	20.49	$n-d+1=9$
	opt.	26.60	$(d-1)/n=1/2$	22.30	$(d-1)/n=1/3$	20.49	$(d-1)/n=1/4$
	Gau.	26.62		22.31		20.50	
	t	26.52		22.27		20.45	
8	*qdf.*	25.18	$n-d+1=7$	21.42	$n-d+1=14$	19.80	$n-d+1=21$
	opt.	25.18	$(d-1)/n=1/2$	21.42	$(d-1)/n=1/3$	19.80	$(d-1)/n=1/4$
	Gau.	25.18		21.42		19.81	
	t	25.12		21.39		19.81	

Table 5: Mean error rate (%) v.s. dimensionality in multivariate two-class problem with common sample covariance matrixes and increasing sample size proportional to dimensionality.

d \ n		$n=d$		$n=2d$		$n=3d$	
2	*qdf.*	30.48	$n-d+1=1$ $(d-1)/n=1/2$	22.51	$n-d+1=3$ $(d-1)/n=1/4$	20.18	$n-d+1=5$ $(d-1)/n=1/6$
	opt.	30.48		22.51		20.18	
	Gau.	30.46		22.50		20.16	
	t	30.41		22.51		20.16	
3	*qdf.*	33.31	$n-d+1=1$ $(d-1)/n=2/3$	22.90	$n-d+1=4$ $(d-1)/n=2/6$	20.32	$n-d+1=7$ $(d-1)/n=2/9$
	opt.	33.31		22.90		20.32	
	Gau.	33.31		22.91		20.33	
	t	33.33		23.00		20.35	
4	*qdf.*	35.23	$n-d+1=1$ $(d-1)/n=3/4$	23.27	$n-d+1=5$ $(d-1)/n=3/8$	20.49	$n-d+1=9$ $(d-1)/n=1/4$
	opt.	35.23		23.27		20.49	
	Gau.	35.25		23.29		20.50	
	t	35.24		23.25		20.45	
8	*qdf.*	39.28	$n-d+1=1$ $(d-1)/n=7/8$	23.69	$n-d+1=9$ $(d-1)/n=7/16$	20.58	$n-d+1=17$ $(d-1)/n=7/24$
	opt.	39.28		23.69		20.58	
	Gau.	39.26		23.70		20.59	
	t	39.18		23.62		20.58	

Table 6: Mean error rate (%) v.s. sample size in 8-variate two-class problem with individual sample covariance matrixes.

n	8	9	10	15	20	30	50	60	80	100
qdf.	8.4484	7.3132	6.3921	4.2008	3.3950	2.7538	2.3265	2.2316	2.1124	2.0433
opt.	7.5237	6.6138	5.8786	4.0325	3.3077	2.7205	2.3156	2.2243	2.1075	2.0404
Keehn	7.5265	6.6176	5.8828	4.0347	3.3099	2.7209	2.3157	2.2243	2.1076	2.0406

Since only the first elements of the mean vectors are different, the reduction of mean error rate with increasing dimensionality is owing to the increase of the degrees of freedom of the t-distribution, i.e. the improvement of the estimation accuracy of the covariance matrix with increasing sample size. Because the sample size and the sample covariance matrixes are common to classes, the optimum classifier and the quadratic classifier give the same results. The mean error rate predicted by the t-quadrature is well coincident to those by Monte Carlo simulation, and the Gaussian quadrature.

Table 5 shows the results when the sample size n increases proportional to the dimensionality d. In this case, $(d-1)/n$ increases with increase of the dimensionality d. The degrees of freedom of t-distribution is fixed when $n=d$, and increases when $n=2d$ or $n=3d$. Under these conditions, the mean error rate increases with the increase of the dimensionality. In this example, the shrinkage of the apparent Mahalanobis distance dominantly affects to and increases the mean error rate as dimensionality increases.

6.1.3 Multivariate case with different sample covariance

Table 6 shows the mean error rates of the optimum classifier and the quadratic classifier for two classes with different sample covariance matrixes. The mean error rates were evaluated by Monte Carlo simulation employing the Bayesian sampling, where the size of test sample and the number of iteration are 5000. The size of design sample and the *a priori* probabilities are common to the classes. The sample covariance matrix of class1 is 8×8 identity matrix, and the one of class2 is 8×8 diagonal matrix with diagonal elements

$$\mathrm{diag}\Sigma_2 = (8.41, 12.06, 0.12, 0.22, 1.49, 1.77, 0.35, 2.73)\ . \tag{66}$$

The mean vectors are those given by (65), and the used optimum discriminant function is (18). The

mean error rates of the quadratic classifier approach to those of the optimum classifier as the sample size n increases, however the optimum classifier outperforms the quadratic classifier for all sample size. Table 6 contains the mean error rates of Keehn's original classifier with $n-1$ degrees of freedom of the Wishart distribution as the population covariance matrixes. It is shown that the Keehn's classifier yields slightly higher mean error rates than the optimum classifier, and that the degrees of freedom of the Wishart distribution should be n when the population mean vector is known.

7 Conclusion

This paper dealt with limited sample based optimum classifier design and the theoretical evaluation of the mean error rate. To verify the optimality of the classifier and the correctness of the mean error calculation, the results of Monte Carlo simulation employing the Bayesian sampling were shown. In the Monte Carlo simulation, the property of the optimum classifier was studied when n_0 was set to zero and the prior distribution was completely neglected. When n_0 is not zero, the mean error rate is expressed by (43) and is further minimized by selecting optimum n_0 which maximizes

$$f(n_0) = \left(1 - \frac{d-1}{n+n_0}\right)(M_2 - M_1)^t \left\{\frac{n}{n+n_0}\Sigma + \frac{n_0}{n+n_0}\Sigma_0\right\}^{-1}(M_2 - M_1)\ . \tag{67}$$

When no prior knowledge about the initial covariance matrix Σ_0 is available, a spherical Gaussian distribution with

$$\begin{aligned} \Sigma_0 &= \sigma^2 I \\ \sigma^2 &= \frac{1}{d}\mathrm{tr}\Sigma \end{aligned} \tag{68}$$

may be assumed, and the optimum n_0 maximizing (67) is determined by numeric search. The apparent Mahalanobis distance in (67) is calculated using a kind of Bayes estimate consisting of a linear combination of the sample covariance matrix and the initial covariance matrix. The Bayes estimate is identical to Σ for $n_0 = 0$, and is identical to Σ_0 for $n_0 = \infty$. Meanwhile, the increase of n_0 has similar effect as the increase of the sample size to add the degrees of freedom of the t-distribution, and to reduce the shrinkage of the apparent Mahalanobis distance. Therefore complete ignorance of the prior distribution by setting n_0 to zero does not lead the best possible classifier.

In practical classifier design, the size of available sample is always limited. As a result the increase of the dimensionality d accelerates the shrinkage of the apparent Mahalanobis distance and the error rate turns to increase at the optimum dimensionality. This phenomenon is known as peaking phenomenon. There had been a long argument on the peaking phenomenon of the optimum classifier. Because the set of all d-variate classifiers does include $(d-1)$-variate optimum classifiers, the optimum classifiers are intuitively expected never to give rise to the peaking phenomenon. However the result of the theoretical analysis by Hughes indicated the peaking phenomenon of the optimum classifier [8]. After a long argument, Campenhaut introduced the concept of comparability and showed that the optimum and comparable classifiers does not yield the peaking phenomenon [11]. The comparability requires the consistent assumption on the prior distributions. The reason why the error rate of the optimum classifiers with $n_0 = 0$ increases as the dimensionality goes up is that they are not comparable.

It is empirically known that the use of the Bayes estimate or a regularized estimate of the covariance matrix instead of the maximum likelihood estimate has an effect to avoid the peaking phenomenon of the quadratic classifier [1, p. 68], [13, 14, 15]. It is also known that the quadratic classifier employing the Bayes estimate leads to practically important classification techniques such as the subspace method [16]. In order to formalize a general principle for designing robust classifiers, further theoretical study on the role of the prior distributions is essentially important.

It is challenging and interesting future research theme to analyze the mean error rate for a multivariate case with different sample covariance matrixes. It is known by theoretical analysis that the effect of limited sample is more severe when the covariance matrixes are not common [5, 6, 17]. However these analyses deal with the quadratic classifier, which is not optimum, and the exact evaluation of the mean error rate is very complicated. There is a possibility that mean error rate of the optimum classifier is expressed in much simpler form as it is in the common covariance case. Taking expectation of the optimum classifier performance over population parameters may generally leads simpler and more meaningful result than conventional approach which takes expectation of non optimum classifier performance over sample parameters.

Theoretical study on the Bayesian sampling and its application to experimental studies are other future research themes. The mixture distribution of the samples drawn by the Bayesian sampling is expected to be converted to the multivariate elliptical t-distribution by (15) or (49), however it is not mathematically proved yet. In most of the real world application, given sample parameters are fixed and the population parameters are unknown. The Bayesian sampling agrees better with these realities than non-Bayesian sampling, and provides us a new way of the Monte Carlo simulation such as the analysis of multi-category classification problems beginning with real world sample parameters at hand.

References

[1] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[2] D. G. Keehn. A note on learning for gaussian properties. *IEEE Trans. Inform. Theory*, Vol. IT-11, No. 1, pp. 126–231, jan 1965.

[3] F. Kimura and M. Shridhar. Handwritten numeral recognition based on multiple algorithms. *Pattern Recognition*, Vol. 24, No. 10, pp. 969–983, oct 1991.

[4] R. Sitgreaves. Some results on the distribution of the w-classification statistic. In *Studies in Item Analysis and Prediction*, pp. 241–261, Stanford, 1961. CA: Stanford Univ. Press.

[5] F. Kimura and M. Shridhar. On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. PAMI*, Vol. PAMI-2, No. 3, pp. 242–252, may 1980.

[6] K. Fukunaga and R. R. Hayes. Effects of sample size in classifier design. *IEEE Trans. PAMI*, Vol. 11, No. 8, pp. 873–885, aug 1989.

[7] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition.* Academic Press, New York, 1990.

[8] G. F. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Inform. Theory*, Vol. IT-14, No. 1, pp. 55–63, jan 1968.

[9] B. Chandrasekaran. Independence of measurements and the mean recognition accuracy. *IEEE Trans. Inform. Theory*, Vol. IT-17, No. 4, pp. 452–456, jul 1971.

[10] W. G. Waller and A. K. Jain. On the monotonicity of the performance of bayesian classifiers. *IEEE Trans. Inform. Theory*, Vol. IT-24, pp. 392–394, 1978.

[11] J. M. Van Campenhout. On the peaking of hughes mean recognition accuracy: The resolution of an apparent paradox. *IEEE Trans. Syst., Man, Cybern.*, Vol. SMC-8, No. 5, pp. 390–395, may 1978.

[12] R. J. Muirhead. *Aspects of Multivariate Statistical Theory.* Wiley, New York, 1982.

[13] S. Tsuruoka F. Kimura, K. Takashina and Y. Miyake. Modified quadratic discriminant functions and the application to chinese character recognition. *IEEE Trans. PAMI*, Vol. PAMI-9, No. 1, pp. 149–153, jan 1987.

[14] J. H. Friedman. Regularized discriminant analysis. *Journal of American Statistical Association,*, Vol. 84, No. 405, pp. 165–175, 1989.

[15] F. Kimura T. Wakabayashi, S. Tsuruoka and Y. Miyake. Study on feature selection in handwritten numeral recognition. *Trans. IEICE, Japan*, Vol. J78-D-II, No. 11, pp. 1627–1638, nov 1995.

[16] M. Shridhar F. Kimura, Y. Miyake. Relationship among quadratic discriminant functions for pattern recognition. In *Proc. of 4th IWFHR*, pp. 418–422, dec 1994.

[17] S. J. Raudys and A.K.Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. PAMI*, Vol. 13, No. 3, pp. 252–264, mar 1991.