The Optimum Classifier and the Performance Evaluation by Bayesian Approach

Xuexian Han, Tetsushi Wakabayashi, and Fumitaka Kimura

Faculty of Engineering, Mie University
1515 Kamihama, Tsu 514-8507, JAPAN

Abstract. This paper deals with the optimum classifier and the performance evaluation by the Bayesian approach. Gaussian population with unknown parameters is assumed. The conditional density given a limited sample of the population has a relationship to the multivariate t-distribution. 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. It is also shown that the Bayesian formulas of the mean error rate have the following characteristics. 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 and the mean error rate is explicitly expressed.

1 Introduction

The Bayesian approach deals with unknown parameters as random variables and assumes their a priori distributions. The essential role of the a priori distribution has not been well known, and the validity of the Bayesian approach and its application has been long argued [1]. 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 is the essential attractiveness of this approach.

This paper deals with the optimum classifier and the performance evaluation by the Bayesian approach. Gaussian population with unknown parameters is assumed. The conditional density given a limited sample of the population has a relationship to the multivariate t-distribution. As a result, the obtained optimum classifier is different from the 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.

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 (not presented in this paper become of the page

F.J. Ferri et al. (Eds.): SSPR&SPR 2000, LNCS 1876, pp. 591–600, 2000.

limit). 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 Bayesian mean error rate has its own interpretation and significance different from those of non-Bayesian 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.

The optimum classifier based on the Bayesian approach 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.

In subsequent sections, a case with unknown covariance matrix (with known mean vector) is described in Section 2 to 4. A new sampling procedure and the result of Monte Carlo simulation are 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.The density $p(X|K)$ is the d-variate Gaussian distribution, and the density $p(K|\chi)$ is the Wishart distribution of n_n degrees of freedom [2,5].

Performing the integration (1), we have

$$\begin{aligned} p(X|\chi) &= (n_n\pi)^{-\frac{d}{2}}|\Sigma_n|^{-\frac{1}{2}}\frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n-d+1}{2}\right)}\left\{1+\frac{1}{n_n}(X-M)^t\Sigma_n^{-1}(X-M)\right\}^{-\frac{n_n+1}{2}} \\ \Sigma_n &= \frac{n_0\Sigma_0+n\Sigma}{n_0+n} \\ \Sigma &= \frac{1}{n}\sum_{i=1}^{n}(X_i-M)(X_i-M)^t \\ n_n &= n_0+n\ , \end{aligned} \tag{2}$$

where M is the population mean vector, and 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. The notation $\Gamma(x)$ denotes the gamma function.

By variable transformation

$$X - M = \sqrt{\frac{n_n}{n_n - d + 1}} T. \tag{3}$$

T leads to the multivariate elliptical t-distribution with $n_n - d + 1$ degrees of freedom [6].

3 Optimum Discriminant Function

The optimum discriminant function for general case is derived from (2) as

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

$$D = (n_n \pi)^{-\frac{d}{2}} \frac{\Gamma\left(\frac{n_n+1}{2}\right)}{\Gamma\left(\frac{n_n-d+1}{2}\right)} \ . \tag{4}$$

4 Evaluation of Mean Error Rate

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

$$\begin{aligned} h(X) &= (M_2 - M_1)^t \Sigma_n^{-1} \sqrt{\frac{n_n - d + 1}{n_n}} X \\ &\quad + \frac{1}{2}\sqrt{\frac{n_n - d + 1}{n_n}} (M_1^t \Sigma_n^{-1} M_1 - M_2^t \Sigma_n^{-1} M_2). \end{aligned} \tag{5}$$

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.The means of $h(X)$ are given by

$$\begin{aligned} \eta_1 &= -\frac{1}{2}\sqrt{\frac{n_n - d + 1}{n_n}} \delta_n^2 \\ \eta_2 &= \frac{1}{2}\sqrt{\frac{n_n - d + 1}{n_n}} \delta_n^2 \\ \delta_n^2 &= (M_2 - M_1)^t \Sigma_n^{-1} (M_2 - M_1) \ . \end{aligned} \tag{6}$$

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

$$\begin{aligned}\sigma_1^2 &= (M_2 - M_1)^t \Sigma_n^{-1} E\{\frac{n_n - d + 1}{n_n}(X - M_1)(X - M_1)^t|\omega_1\}\Sigma_n^{-1}(M_2 - M_1) \\ &= \frac{n_n - d + 1}{n_n - d - 1}(M_2 - M_1)^t \Sigma_n^{-1}(M_2 - M_1) = \frac{n_n - d + 1}{n_n - d - 1}\delta_n^2 \\ \sigma_2^2 &= \frac{n_n - d + 1}{n_n - d - 1}\delta_n^2 \ . \end{aligned} \tag{7}$$

Using these parameters the mean error rate is given by

$$\begin{aligned}\varepsilon &= P(\omega_1)\varepsilon_1 + P(\omega_2)\varepsilon_2 \\ &= \frac{1}{2}\Phi_{n_n - d+1}\left(\frac{\eta_1}{\delta_n}\right) + \frac{1}{2}\left\{1 - \Phi_{n_n - d+1}\left(\frac{\eta_2}{\delta_n}\right)\right\} \\ &= 1 - \Phi_{n_n - d+1}\left(\frac{1}{2}\sqrt{\left(1 - \frac{d-1}{n_n}\right)\delta_n^2}\right) \ . \end{aligned} \tag{8}$$

When $n_0 = 0$,

$$\varepsilon = 1 - \Phi_{n-d+1}\left(\frac{1}{2}\sqrt{\left(1 - \frac{d-1}{n}\right)\delta^2}\right) \ . \tag{9}$$

The function $\Phi_n(x_0)$is defined by

$$\Phi_n(x_0) = \int_{-\infty}^{x_0} t_n(x)dx \ , \tag{10}$$

where t_n is the univariate t-disrtibution with n degrees of freedom.

The Bayesian formulas of the mean error rate (8) and (9) have the following characteristics when compared with the non-Bayesian formulas.

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 (9) is an apparent one which is calculated using the known population mean vector and the sample covariance matrix. (9) 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 (Fig. 1). 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 (3), and causes a problem so called "peaking phenomenon" or "curse of dimensionality"[3, 4,7]. 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 6.

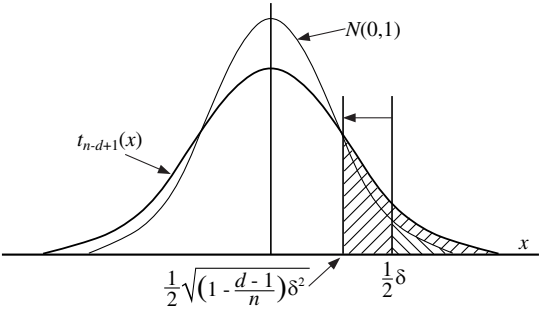

Fig. 1. Increase of mean error rate due to limited sample effect

5 Computer Simulation

5.1 Bayesian Sampling

In the following computer simulation, a new sampling procedure called Bayesian sampling is employed together with the ordinary sampling procedure. Fig. 2 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. Fig. 2 (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 for designing 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 for 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.

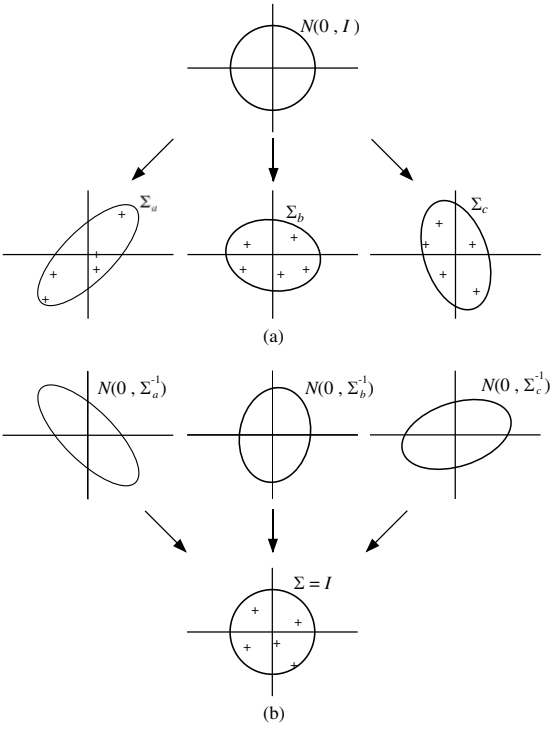

Fig. 2. Relationship between ordinary sampling (a), and Bayesian sampling (b)

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.

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\Sigma_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu_1)(X_i - \mu_1)^t \ . \tag{11}$$

By setting

$$Y = \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t (X - \mu_1) \qquad (\Sigma_1 \Phi_1 = \Phi_1 \Lambda_1) \tag{12}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = 0$$
$$\frac{1}{n-1}\sum_{i=1}^{n} Y_i Y_i^t = \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \Sigma_1 \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t = \Phi_1 \Phi_1^t = I \tag{13}$$

and the population parameters of Y are given by

$$E(Y) = -\Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \mu_1$$
$$V(Y) = E\left[\{Y - E(Y)\}\{Y - E(Y)\}^t\right] = \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t = \Sigma_1^{-1} , \quad (14)$$

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

Further by setting

$$\begin{aligned} Z &= \Phi_2 \Lambda_2^{\frac{1}{2}} Y + \mu_2 \\ &= \Phi_2 \Lambda_2^{\frac{1}{2}} \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t X + \mu_2 - \Phi_2 \Lambda_2^{\frac{1}{2}} \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \mu_1 \quad (\Sigma_2 \Phi_2 = \Phi_2 \Lambda_2) \end{aligned} \quad (15)$$

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

$$\frac{1}{n} \sum_{i=1}^{n} Z_i = \Phi_2 \Lambda_2^{\frac{1}{2}} \frac{1}{n} \sum_{i=1}^{n} Y_i + \mu_2 = \mu_2$$
$$\frac{1}{n-1} \sum_{i=1}^{n} Y_i Y_i{}^t = \Phi_2 \Lambda_2^{\frac{1}{2}} I \Lambda_2^{\frac{1}{2}} \Phi_2^t = \Phi_2 \Lambda_2 \Phi_2^t = \Sigma_2 \quad (16)$$

and the population parameters of Z are given by

$$E(Z) = \mu_2 - \Phi_2 \Lambda_2^{\frac{1}{2}} \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \mu_1$$
$$V(Z) = \Phi_2 \Lambda_2^{\frac{1}{2}} \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \Phi_1 \Lambda_1^{-\frac{1}{2}} \Phi_1^t \Lambda_2^{\frac{1}{2}} \Phi_2^t = \Phi_2 \Lambda_2^{\frac{1}{2}} \Sigma_1^{-1} \Lambda_2^{\frac{1}{2}} \Phi_2^t . \quad (17)$$

When the population mean vector M is known, (17) is replaced by

$$E(Z) = M$$
$$V(Z) = \Phi_2 \Lambda_2^{\frac{1}{2}} \Sigma_1^{-1} \Lambda_2^{\frac{1}{2}} \Phi_2^t$$
$$\Sigma_1 = \frac{1}{n} \sum_{i=1}^{n} (X_i - M)(X_i - M)^t. \quad (18)$$

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

Multivariate Case with Common Sample Covariance Matrix. Table. 1 and Fig. 3 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 sim. 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 row t shows the mean error rate by (9).

The optimum discriminant function employed in the simulation is derived from (4). The sample covariance matrix is $d \times d$ identity matrix, and the population mean vectors are

$$M_1 = (0, 0, 0, \cdots, 0),$$
$$M_2 = (1, 1, 1, \cdots, 1). \quad (19)$$

Table 1. Mean error rate (%) v.s. dimensionality in multivariate two-class problem with common sample covariance matrixes

	n	10	15	20
d		*opt.*(*qdf.*)	*opt.*(*qdf.*)	*opt.*(*qdf.*)
2	*sim.*	25.97	25.29	24.95
	t	25.96	25.28	24.95
4	*sim.*	21.49	19.44	18.45
	t	21.52	19.43	18.47
6	*sim.*	21.26	16.94	15.17
	t	21.30	17.04	15.28
8	*sim.*	24.65	16.49	13.58
	t	24.75	16.59	13.74
10	*sim.*	35.32	17.65	13.22
	t	35.24	17.80	13.29

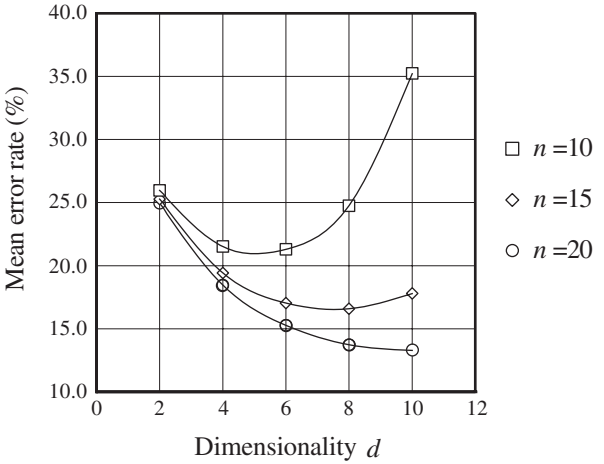

Fig. 3. Theoretical mean error rate (%) v.s. dimensionality

For these parameters, the Mahalanobis distance $\delta^2 = n$ and (9) is minimized when $d = (n+1)/2$.

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 rates predicted by the t-distribution is well coincident to those by Monte Carlo simulation.

Multivariate Case with Different Sample Covariance. Fig. 4 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

$$\mathrm{diag}\Sigma_2 = (8.41, 12.06, 0.12, 0.22, 1.49, 1.77, 0.35, 2.73)\ . \tag{20}$$

The mean vectors are given by

$$\begin{aligned} M_1 &= (-1, 0, 0, \cdots, 0), \\ M_2 &= -M_1. \end{aligned} \tag{21}$$

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.

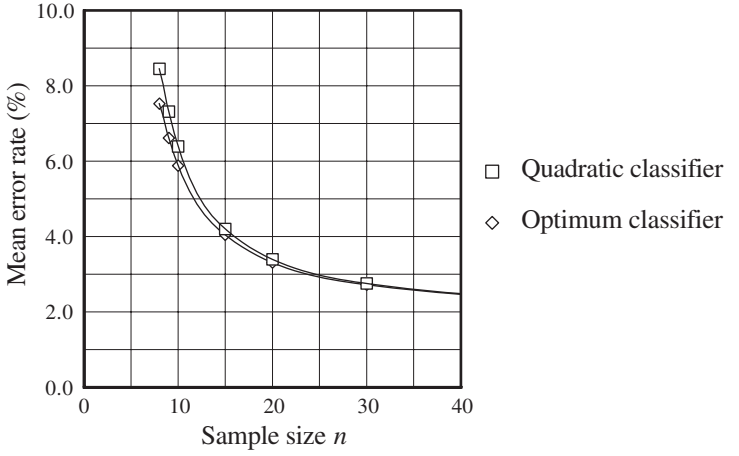

Fig. 4. Mean error rate of quadratic classifier and optiomum classfier v.s. sample size in 8-variate two-class problem with individual sample covariance matrixes

6 Conclusion and Discussion

¡This paper dealt with the optimum classifier design and the performance evaluation by the Bayesian aproach. 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. It was also shown that the Bayesian formulas of the mean error rate have the following characteristics.

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 and the mean error rate is explicitly expressed.

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 (8) and is further minimized by selecting optimum n_0 which maximizes

$$f(n_0) = \left(1 - \frac{d-1}{n+n_0}\right)(M_2 - M_1)^t \left\{\frac{n}{n+n_0}\Sigma + \frac{n_0}{n+n_0}\Sigma_0\right\}^{-1}(M_2 - M_1)\ . \tag{22}$$

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 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. New York: Wiley, 1973, p.76, p.68.
2. D. G. Keehn, "A Note on Learning for Gaussian Properties," IEEE Trans. Inform. Theory, vol. IT-11, no. 1, pp.126-132, Jan 1965.
3. S. Raudys and V. Pikelis, "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.
4. 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.
5. K. Fukunaga, Introduction to Statistical Pattern Recognition, Second Edition, New York: Academic Press, 1990, pp.392-393, pp.91-92.
6. R. J. Muirhead, Aspects of Multivariate Statistical Theory. New York: Wiley, 1982, p.32-49.
7. 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.