The Annals of Statistics
1981, Vol. 9, No. 5, 1002–1009

ASYMPTOTIC EXPANSIONS FOR CORRECT CLASSIFICATION RATES IN DISCRIMINANT ANALYSIS

BY MARK J. SCHERVISH

Carnegie-Mellon University

When classifying an observation into one of k multivariate normal distributions based on samples of correctly classified observations, two estimates of the probability of correct classification, called the apparent and plug-in correct classification rates, are considered. Asymptotic expansions are found for the means and variances of these estimates. It is shown that these expansions can be used to help reduce the bias of the estimates. In the course of finding the expansions, an asymptotic expansion for the conditional joint density of two observations given the sample mean and pooled covariance matrix is found.

1. Introduction. Much work has been done concerning the probability of correctly classifying a random vector X as belonging to one of two multivariate normal populations, c.f. Lachenbruch (1975). Suppose now that X has one of $k \geq 2$ p-variate normal distributions, denoted by $P_i = N_p(\mu_i, V)$ with V nonsingular. For a simple loss function, Rao (1954) finds the Bayes classification rule with respect to the prior distribution which assigns probability π_i to distribution P_i. The Bayes rule simplifies to:

Classify X as population j if, for $i = 1, \cdots, k$,

$$\log(\pi_i/\pi_j) + (\mu_i - \mu_j)'V^{-1}X + \tfrac{1}{2}(\mu_j'V^{-1}\mu_j - \mu_i'V^{-1}\mu_i) \leq 0. \tag{1.1}$$

If the parameters are unknown and a sample $X_{i1}, \cdots, X_{in_i}$ from P_i is available for $i = 1, \cdots, k$, then the Bayes rule (1.1) can be estimated by using the estimate $T = (\bar{X}_1, \cdots, \bar{X}_k, W)$ of $(\mu_1, \cdots, \mu_k, V)$, where W is the pooled sum of products matrix divided by its degrees of freedom n.

There is no loss of generality in considering only the correct classification rate for population (distribution) k, since the populations can be renumbered. The estimate of rule (1.1) is called the *sample rule* and for $j = k$ is given by:

Classify X as population k if for $i = 1, \cdots, \ell$,

$$\log(\pi_i/\pi_k) + (\bar{X}_i - \bar{X}_k)'W^{-1}X + \tfrac{1}{2}(\bar{X}_k'W^{-1}\bar{X}_k - \bar{X}_i'W^{-1}\bar{X}_i) \leq 0, \tag{1.2}$$

where $\ell = k - 1$. The distribution of the statistics (1.2) is the same for all nonsingular V, hence there is no loss of generality in assuming

$$V = I. \tag{1.3}$$

Following Lachenbruch (1975, page 30), define the *apparent correct classification rate* for population i, denoted by ACCR_i, to be the proportion of $X_{i1}, \cdots, X_{in_i}$ which are classified correctly by the sample rule. Another estimate of correct classification rate is obtained by pretending that the estimates T are the parameters and computing the probability that (1.2) occurs when $X \sim P_k$. This estimate will be called the *plug-in correct classification rate*, PCCR_k. The purpose of this article is to find asymptotic expansions for the means and variances of ACCR_k and PCCR_k with errors which are $O(N^{-2})$, where $N = \min\{n_1, \cdots, n_k, n\}$. The leading term $O(1)$ of the mean expansions will be the probability that the Bayes rule (1.1) classifies X correctly when $X \sim P_k$. These expansions will then be used to find improved estimators of the error rate.

Received November 30, 1979; revised February 2, 1981
AMS 1970 *subject classifications.* Primary 62H30; secondary 62E20.
Key words and phrases. Classification, error rates, plug in error rate, apparent error rate.

It has been noted by Lachenbruch (1975, page 33), among others, that the apparent error rate ($1 - \text{ACCR}_t$) is not a particularly good estimator of error rate. It is noted in Section 6 that with the aid of the asymptotic expansions the bias of 1-ACCR_t as well as the bias of 1-PCCR_t as estimators of the error rate for the sample rule can be reduced. In Section 6, the results of Monte Carlo studies are given in which the bias-corrected ACCR_t and PCCR_t are compared to other estimators of error rate. McLachlan (1973, 1976) considers ACCR and PCCR in the special case $k = 2$. The present work, in addition to handling the case of any k, attempts to make more clear the connection between the asymptotic expansions for ACCR and PCCR by deriving them simultaneously. The results will be stated without proofs in the text. The proofs are given in the appendix for the serious reader since they are mostly quite technical.

2. The conditional distribution of two observations given the sample moments. The following assumption is necessary for the remainder of this work.

ASSUMPTION 1. *The vectors $\mu_1, \cdots, \mu_k$ do not lie in any flat of dimension $k - 2$ or less.*

Under Assumption 1, the dimension p is greater than or equal to $\ell = k - 1$. This fact is needed in Lemma 1 below as well as in Theorems 1 and 2.

Write ACCR_k as

$$\text{ACCR}_k = n_k^{-1} \sum_{i=1}^{n_k} I_i,$$

where I_i equals one if X_{ki} is classified correctly by the sample rule (1.2) and equals zero if not. It is clear that $E(\text{ACCR}_k) = E(I_1)$, and that $E(I_1)$ can be written as $EE(I_1 \mid T)$ where T is the sample means and covariance matrix. Similarly $E(\text{ACCR}_k^2) = n_k^{-1}E(I_1) + (n_k - 1)n_k^{-1}E(I_1I_2)$, and $E(I_1I_2) = EE(I_1I_2 \mid T)$. Since I_1 and I_2 are measurable functions of X_{k1}, X_{k2} and T, it would help to know the conditional joint distribution of (X_{k1}, X_{k2}) given T. Lemma 1 is useful in this regard.

LEMMA 1. *Under Assumption* 1 *and condition* (1.3), *the conditional joint density of* (X_{k1}, X_{k2}) *given* $T = (\bar{X}_1, \cdots, \bar{X}_k, W)$ *is*

$$\begin{aligned} f(x_1, x_2) &= (2\pi)^{-p} \mid W \mid^{-1} \exp\{-\tfrac{1}{2}(A_{11} + A_{22})\} \\ &\cdot [1 + pn_k^{-1} - (2n)^{-1}p(p+3) - n_k^{-1}A_{12} \\ &+ \{(p+3)(2n)^{-1} - (2n_k)^{-1}\}(A_{11} + A_{22}) - (2n)^{-1}A_{12}^2 \\ &- (4n)^{-1}(A_{11}^2 + A_{22}^2)] + G, \end{aligned} \tag{2.1}$$

where

$$A_{ij} = (x_{ki} - \bar{X}_k)'W^{-1}(x_{kj} - \bar{X}_k), \; ij = 1, 2,$$

and G satisfies

$$E \int_{\Omega} G \, dx_{k1} dx_{k2} = O(N^{-2}), \tag{2.2}$$

for any subset Ω of $\mathbb{R}^p \times \mathbb{R}^p$.

To find the mean and variance of ACCR_k, the term G is ignored, the remainder of f is integrated over the appropriate subsets of $\mathbb{R}^p \times \mathbb{R}^p$, and the expected value is taken over the distribution of T. If one were to ignore all of the terms in (2.1) enclosed in square brackets, what would remain would be the density of two independent random variables each with $N_p(\bar{X}_k, W)$ distribution. But this is exactly the distribution one assumes for X when computing PCCR_k as the probability that (1.2) occurs. It is not surprising, therefore,

that the asymptotic expansions for the means and variances of ACCR_k and PCCR_k are so similar.

3. The expected value expansions. First define $\sigma_{ij} = (\mu_i - \mu_k)'(\mu_j - \mu_k)$ and $\Sigma = ((\sigma_{ij}))$. Then set $S = ((S_{ij})) = HH'$, where

$$H = \begin{bmatrix} (\bar{X}_1 - \bar{X}_k)'W^{-1/2} \\ \cdot \\ \cdot \\ \cdot \\ (\bar{X}_\ell - \bar{X}_k)'W^{-1/2} \end{bmatrix}$$

In view of (1.2), define

$$Y_i = HW^{-1/2}(X_{ki} - \bar{X}_k) + \varepsilon, \quad i = 1, 2, \tag{3.1}$$

where

$$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_\ell)' \text{ with } \varepsilon_i = \tfrac{1}{2}\sigma_{ii} - \tfrac{1}{2}S_{ii}.$$

If we define

$$B = \{y \in \mathbb{R}^\ell : y_i \leq \tfrac{1}{2}\sigma_{ii} - \log(\pi_i/\pi_k),\ i = 1, \cdots, \ell\},$$

then $E(I_1 \mid T) = P(Y_1 \in B \mid T)$ and $E(I_1 I_2 \mid T) = P(Y_1 \in B,\ Y_2 \in B \mid T)$. For the mean of ACCR_k, we need only $E(I_1 \mid T)$, so we compute the conditional marginal density of Y_1 given T using the transformation (3.1) on the density (2.1). The result is, ignoring higher order terms,

$$\begin{aligned} g(y) = h^*(y)[1 &+ \ell(2n_k)^{-1} - (2n_k)^{-1}(y - \varepsilon)'S^{-1}(y - \varepsilon) - \ell(\ell + 2)(4n)^{-1} \\ &+ (\ell + 2)(2n)^{-1}(y - \varepsilon)'S^{-1}(y - \varepsilon) - (4n)^{-1}\{(y - \varepsilon)'S^{-1}(y - \varepsilon)\}^2], \end{aligned} \tag{3.2}$$

where

$$h^*(y) = (2\pi)^{-\ell/2} \mid S \mid^{-1/2} \exp\{-\tfrac{1}{2}(y - \varepsilon)'S^{-1}(y - \varepsilon)\}.$$

The mean of ACCR_k will then equal $E\int_B g(y)dy + O(N^{-2})$. It is not hard to see that if ρ_k denotes the probability that the Bayes rule (1.1) classifies X correctly when $X \sim P_k$, then

$$\rho_k = \int_B \phi(y)\, dy, \tag{3.3}$$

with $\phi(y) = (2\pi)^{-\ell/2} \mid \Sigma \mid^{-1/2} \exp(-\tfrac{1}{2}\, y'\Sigma^{-1}y)$. Substituting the estimates T for the parameters $(\mu_1, \cdots, \mu_k, V)$ into (3.3) it follows that the plug-in correct classification rate,

$$\mathrm{PCCR}_k = \int_B h^*(y)\, dy. \tag{3.4}$$

What we see here is that since the conditional distribution of X_{k1} given T is "nearly" $N_p(\bar{X}_k, W)$, $E(\mathrm{ACCR}_k \mid T)$ is "nearly" equal to PCCR_k, the difference, to order N^{-1}, being the integral over B of $h^*(y)$ times the $O(N^{-1})$ terms which appear in (3.2).

THEOREM 1: *Under Assumption* 1 *and condition* (1.3),

$$E(\mathrm{ACCR}_k) = \rho_k + R_1 + R_2 + R_3 + R_4 + O(N^{-2}),$$

$$E(\mathrm{PCCR}_k) = \rho_k + R_1 + R_2^* + R_3 + R_4^* + O(N^{-2}),$$

where ρ_k *is defined in* (3.3)

$$R_1 = \sum C_i^1 J_i^1, \qquad R_2 = \sum\sum C_{i,j}^2 J_{i,j}^2, \qquad R_2^* = \sum\sum D_{i,j}^2 J_{i,j}^2,$$

$$R_3 = \sum\sum\sum C^3_{i,j,q} J^3_{i,j,q}, \qquad R_4 = \sum\sum\sum\sum C^4_{i,j,q,t} J^4_{i,j,q,t}, \qquad R^*_4 = \sum\sum\sum\sum D^4_{i,j,q,t} J^4_{i,j,q,t},$$

with all summations being over integers 1, 2, $\cdots$, $\ell = k - 1$ *and*

$$C^1_i = -p(2n)^{-1}\sigma_{ii} + (4-p)(2n_i)^{-1} + (4 - p + 2\ell)(2n_k)^{-1} + n^{-1}\sum_{r=1}^{\ell}\sigma_{ir} + \sum_{r=1}^{\ell} n_r^{-1}\sigma^{rr}\sigma_{ir},$$

$$C^2_{i,j} = (4n)^{-1}\sigma^2_{ij} + (2n)^{-1}\sigma_{ij}(1+p) - (2n_k)^{-1}(\ell - p + 4 + \sigma_{ij}\sum_{r=1}^{\ell}\sum_{q=1}^{\ell}\sigma^{rq})$$
$$+ (2n_i)^{-1}\delta_{ij}(\sigma_{ij} - 4 - \ell + p) - \tfrac{1}{2}\sigma_{ij}\sum_{q=1}^{\ell}\sigma^{qq}n_q^{-1},$$

$$C^3_{i,j,q} = -(2n)^{-1}\sigma_{iq}\sigma_{ij} - (2n_k)^{-1}(\sigma_{iq} + \sigma_{ij}) - (2n_i)^{-1}\delta_{ij}\sigma_{iq} - (2n_i)^{-1}\delta_{iq}\sigma_{ij}$$

$$C^4_{i,j,q,t} = -(4n)^{-1}\sigma_{ij}\sigma_{qt} + (8n)^{-1}(\sigma_{iq}\sigma_{jt} + \sigma_{it}\sigma_{jq}) + (8n_j)^{-1}\delta_{jt}\sigma_{iq}$$
$$+ (8n_j)^{-1}\delta_{jq}\sigma_{it} + (8n_i)^{-1}\delta_{it}\sigma_{jq} + (8n_i)^{-1}\delta_{iq}\sigma_{jt} + (8n_k)^{-1}(\sigma_{iq} + \sigma_{jt} + \sigma_{it} + \sigma_{jq}).$$

$$D^2_{i,j} = C^2_{i,j} - \sigma_{i,j}\{(\ell+2)(2n)^{-1} - (2n_k)^{-1}\}, \quad D^4_{i,j,q,t} = C^4_{i,j,q,t} + (4n)^{-1}\sigma_{ij}\sigma_{qt}.$$

Also,

$$J^1_i = \int_B \sigma^i y \phi(y)\, dy,$$

$$J^2_{i,j} = \int_B (\sigma^i y)(\sigma^j y)\phi(y)\, dy - \sigma^{ij}\rho_k,$$

$$J^3_{i,j,q} = \int_B (\sigma^i y)(\sigma^j y)(\sigma^q y)\phi(y)\, dy,$$

$$J^4_{i,j,q,t} = \int_B (\sigma^i y)(\sigma^j y)(\sigma^q y)(\sigma^t y)\phi(y)\, dy - \rho_k(\sigma^{ij}\sigma^{qt} + \sigma^{iq}\sigma^{jt} + \sigma^{it}\sigma^{jq}),$$

where σ^i *is the* i*th row of* Σ^{-1}, *and* σ^{ij} *is the* (i,j) *element of* Σ^{-1}.

The theorem is stated in a rather structured format because, in the course of developing the expansion, it is discovered that each term has a factor of the form $\int_B P(y)\phi(y)\, dy$, where $P(y)$ is a monomial of degree at most 4 in the coordinates of y. It turns out to be convenient, both for the proof of the theorem and later for the calculation of the expansion via computer, to collect the terms of the expansion by the degree of the monomial $P(y)$. These are the R terms. The J's are the integrals of the monomials, and the C's and D's are the expected values of the coefficients.

Theorem 1 was stated for the special case (1.3), since the general case can be so reduced. For arbitrary nonsingular V, replace μ_i by $V^{-1/2}\mu_i$, $i = 1, \cdots, k$ in all places. In particular, σ_{ij} becomes $(\mu_i - \mu_k)'V^{-1}(\mu_j - \mu_k)$. This remark will apply to the remainder of the results in this article as well as to Theorem 1.

4. The special case $k = 2$. McLachlan (1973, 1976) considered the case of two populations, i.e. $k = 2$, $\ell = 1$. It can be shown that the results of Theorem 1 agree with those of McLachlan for both $E(\text{ACCR}_2)$ and $E(\text{PCCR}_2)$ to terms of order $O(N^{-1})$. Rather than carry out all of the calculations, we will just check that $E(\text{ACCR}_2) - E(\text{PCCR}_2)$ agrees. McLachlan (1976) gives this difference in the case $\pi_1 = \pi_2$ as

$$g(\Delta/32)\{8/n_2 - (12 - \Delta^2)/n\}, \tag{4.1}$$

where $\Delta^2 = \sigma_{11}$ and g is the standard normal density evaluated at $-\tfrac{1}{2}\Delta$. The relevant terms from Theorem 1 are

$$J^2_{11} = -\tfrac{1}{2}\Delta^{-1}g,$$

$$J^4_{1111} = -(1/8\Delta^2 + 3/2)\Delta^{-3}g,$$

$$C^2_{11} - D^2_{11} = \Delta^2(3/n - 1/n_2)/2,$$

$$C^4_{1111} - D^4_{1111} = -\Delta^4(4n).$$

It follows immediately that (4.1) equals $J^2_{11}(C^2_{11} - D^2_{11}) + J^4_{1111}(C^4_{1111} - D^4_{1111})$. The fact that the expansion for $E(\text{PCCR}_2)$ agrees with McLachlan (1973) can also be checked easily.

5. The variance expansions. Theorem 2 gives asymptotic expansions for the variances of ACCR_k and PCCR_k.

THEOREM 2: *Under the conditions of Theorem* 1,

$$\text{Var}(\text{ACCR}_k) = n_k^{-1}\rho_k(1 - \rho_k) + R_5 + R_6 + R_7 + O(N^{-2}),$$

$$\text{Var}(\text{PCCR}_k) = R_5^* + R_6 + R_7^* + O(N^{-2}),$$

where, with summations over $i, \cdots, l = k - 1$,

$$R_5 = \sum\sum J^1_i J^1_j F^2_{i,j}, \qquad R_5^* = \sum\sum J^1_i J^1_j G^2_{i,j}, \qquad R_6 = 2\sum\sum\sum J^1_i J^2_{j,q} C^3_{i,j,q},$$

$$R_7 = 2\sum\sum\sum\sum J^2_{i,j} J^2_{q,t} C^4_{i,j,q,t}, \qquad R_7^* = 2\sum\sum\sum\sum J^2_{i,j} J^2_{q,t} D^4_{i,j,q,t},$$

the J, C, and D terms are as in Theorem 1, *and*

$$F^2_{i,j} = (2n)^{-1}\sigma^2_{ij} + \delta_{ij} n_i^{-1}\sigma_{ij}$$

$$G^2_{i,j} = F^2_{i,j} + n_k^{-1}\sigma_{ij}.$$

Note that the first term of $\text{Var}(\text{ACCR}_k)$ is just the variance of a random variable with binomial $b(n_k, \rho_k)$ distribution. The distribution of ACCR_k would be $b(n_k, \rho_k)$ if the Bayes rule (1.1) were used instead of the sample rule. The terms $R_5 + R_6 + R_7$, then, represent the increase in variance due to using the sample rule.

6. Estimating error rates. The goal of error rate analysis is often to estimate what Hills (1966) calls the actual error rate, which will be denoted $e_k(T)$, i.e. the conditional probability given T that the sample rule (1.2) classifies a future observation X, independent of T, incorrectly if $X \sim P_k$. Both 1-PCCR_k and 1-ACCR_k tend to be optimistic, i.e. low, estimates of $e_k(T)$. Schervish (1981) gives an asymptotic expansion for $E\{1 - e_k(T)\}$ which is similar to those of the present article. If we write that expansion as $E\{1 - e_k(T)\} = \rho_k + Q_1 + O(N^{-2})$, and we write the expansions of the present article as $E(\text{ACCR}_k) = \rho_k + Q_2 + O(N^{-2})$ and $E(\text{PCCR}_k) = \rho_k + Q_3 + O(N^{-2})$, then one can use $e_A = 1 - \text{ACCR}_k + \hat{Q}_2 - \hat{Q}_1$ or $e_p = 1 - \text{PCCR}_k + \hat{Q}_3 - \hat{Q}_1$ to estimate $e_k(T)$, where $\hat{Q}_i$ is the estimate of Q_i formed by using $(\bar{X}_1, \cdots, \bar{X}_k, W)$ as if they were $(\mu_1, \cdots, \mu_k, V)$. The biases of the estimates e_A and e_p should be smaller than those of ACCR_k and PCCR_k as estimates of $e_k(T)$.

Simulations were performed on the Carnegie-Mellon University DEC-20 system to compare the estimates e_A and e_p to 1-ACCR_k, 1-PCCR_k, the "leave one out" estimator of Lachenbruch and Mickey (1968) $(L - M)$, and the "bootstrap" estimator using 100 bootstraps per simulation. This last estimator equals 1-$\text{ACCR}_k + \tilde{R}$ where $\tilde{R}$ is the estimate of $e_k(T) - (1\text{-ACCR}_k)$ described by Efron (1979). In some of the simulations the sample mean vectors nearly violated Assumption 1. In such cases the matrix which estimates Σ is nearly singular and the estimates e_A and/or e_p can become negative or greater than one. When e_A was out of range, it was replaced by ACCR_k. When e_p was out of range, it was replaced by PCCR_k. Let m denote the number of times such replacement was necessary. Two cases were simulated 1000 times each. Each case had $\mu_1' = [1, 0]$, $\mu_2' = [0, 1]$, $\mu_3' = [0, 0]$, and $V = I$. The first case had $n_i = 10$ for each i, and the second had $n_i = 20$. The results are given in Table 1. Using both bias and MSE equal to the average of $\{X - e_3(T)\}^2$ over the 1000 simulations as measures of how close an estimate X is to $e_3(T)$, the results are conflicting. The bootstrap estimate had the smallest bias, but nearly the largest MSE, while e_p had the smallest MSE and the second or third smallest bias. These results are similar to results reported by McLachlan (1980) for the case $k = 2$.

TABLE 1
Simulation results to compare estimates of error rate

Estimator Name	Sample Size 10 Value	MSE	m	Sample Size 20 Value	MSE	m
1-ACCR$_3$	0.4753	0.028		0.5063	0.013	
e_A	0.5382	0.025	15	0.5364	0.013	1
1-PCCR$_3$	0.4837	0.014		0.5131	0.007	
e_P	0.5411	0.014	6	0.5403	0.007	0
L-M	0.4723	0.026		0.5063	0.013	
Bootstrap	0.5435	0.026		0.5397	0.014	
e_3 (T)	0.5462			0.5385		

Each case is based on 1000 simulations. MSE is the average squared deviation of estimator from $e_k(T)$, and m is the number of times e_A or e_P was outside (0,1). For an explanation of the "bootstrap" estimator, see Efron (1979). There were one hundred bootstraps per simulation.

7. Conclusion. It appears that asymptotic expansions for the means of error rate estimators can be used to provide estimates of error rate with less bias and less mean squared error than the original estimators. The new estimators even appear to be competitive with the well known current estimators in terms of bias and mean squared error, when the assumption of normal distributions holds. The J integrals in the theorems can be evaluated using multivariate integration by parts. Computer programs for their evaluation are available in Schervish (1979).

Acknowledgment. Most of this work was performed as part of the author's Ph.D. thesis under Professor R.A. Wijsman at the University of Illinois at Urbana-Champaign.

APPENDIX

PROOF OF LEMMA 1. Define S_2 to be the sum of squares and products matrix for all of the observations except X_{k1} and X_{k2}. Then $nW = S_2 + U_1U_1' + U_2U_2'$ where $U_i = (n_k - i)^{1/2}(n_k - i + 1)^{-1/2}(X_{ki} - \hat{X}_i)$, and $\hat{X}_i = (n_k - i)^{-1}\sum_{j=i+1}^{n_k} X_{kj}$. A theorem of Khatri (1959) and a slight change of variables gives that W is independent of $(Z_1, Z_2) = (W^{-1/2}U_1, W^{-1/2}U_2)$. In fact $(W, \bar{X}_1, \cdots, \bar{X}_k, Z_1, Z_2)$ are jointly independent. The joint density of (Z_1, Z_2), given by Khatri, is

$$g(z_1, z_2) = \begin{cases} c_n |B|^{1/2(n-p-3)} & \text{if } B \text{ is positive definite,} \\ 0 & \text{otherwise,} \end{cases}$$

where

$$c_n = (n\pi)^{-p} \prod_{i=1}^{p} \Gamma\{\tfrac{1}{2}(n - i + 1)\}/\Gamma\{\tfrac{1}{2}(n - i - 1)\},$$

and

$$B = I - n^{-1}(z_1z_1' + z_2z_2').$$

It is not difficult to show that

$$|B| = 1 - n^{-1}[z_1'z_1 + z_2'z_2 - n^{-1}\{(z_1'z_1)(z_2'z_2) - (z_1'z_2)^2\}].$$

It is also true that if $\tau = \{(z_1, z_2): z_i'z_i \leq rn, i = 1, 2\}$, where $0 < r < \frac{1}{2}$, then

$$\int_{\tau^c} g(z_1, z_2)\, dz_1 dz_2$$

is exponentially bounded as n goes to infinity for fixed r. On the set τ, $0 < 1 - |B| < 1$, so

Taylor's theorem can be used to write

$$n \log|B| = -z_1'z_1 - z_2'z_2 - n^{-1}(z_1'z_2)^2 - (2n)^{-1}\{(z_1'z_1)^2 + (z_2'z_2)^2\} + n^{-2}h_n(z),$$

where $h_n(z)$ is a function which behaves like a polynomial for large $z_i'z_i$ for (z_1, z_2) in τ. After approximating c_n and using Taylor's theorem to approximate $\exp(\log|B|)$, we can drop the terms of polynomial order in $z_i'z_i$ which have factors of n^{-2}. Then notice that what remains has an exponentially bounded integral over τ^c. It follows that

$$\begin{aligned} g(z_1, z_2) &= (2\pi)^{-p}\exp\{-\tfrac{1}{2}(z_1'z_1 + z_2'z_2)\}[1 - (2n)^{-1}p(p+3)] \\ &\quad + (p+3)(2n)^{-1}(z_1'z_1 + z_2'z_2) - (2n)^{-1}(z_1'z_2)^2 \\ &\quad -(4n)^{-1}\{(z_1'z_1)^2 + (z_2'z_2)^2\} + H, \end{aligned} \tag{A.1}$$

where $\int_\Omega H\, dz_1\, dz_2 = O(n^{-2})$ for any subset Ω of $\mathbb{R}^p \times \mathbb{R}^p$. Finally, note that $X_{k1} = \bar{X}_k + n_k^{-1/2}(n_k - 1)^{1/2}W^{1/2}Z_1$ and

$$X_{k2} = \bar{X}_k + (n_k - 1)^{-1/2}(n_k - 2)^{1/2}W^{1/2}Z_2 - \{n_k(n_k - 1)\}^{-1/2}W^{1/2}Z_1.$$

Apply this transformation to the density (A.1) to obtain the joint density for (X_{k1}, X_{k2}) given by Lemma 1.

PROOF OF THEOREM 1. The expression $\int_B h^*(y)\, dy$ is of the same form as the actual correct classification rate discussed in Schervish (1981) with ε and S defined slightly differently. The procedure used was to write $h^*(y)$ as $\phi(y)\{1 + Q(y)\}$, where Q is a fourth degree polynomial whose coefficients have means of $O(N^{-1})$ plus higher order terms. The same method applies here. The means of the degree i terms are R_i or R_i' in the expansion for $E(\text{PCCR}_k)$. For full details see Schervish (1979). The difference between $E(\text{ACCR}_k \mid T)$ and PCCR_k is, after taking care to ignore higher order terms whose means are $O(N^{-2})$,

$$\int_B \phi(y)[(2n_k)^{-1}(\ell - y'\Sigma^{-1}y) + (4n)^{-1}\{2y'\Sigma^{-1}y(\ell+2) - \ell(\ell+2) - (y'\Sigma^{-1}y)^2\}]\, dy, \tag{A.2}$$

where h^* has been replaced by ϕ, S^{-1} by Σ^{-1}, and ε by zero since each term already has a factor $O(n^{-1})$ and the differences $S^{-1} - \Sigma^{-1}$, $h^* - \phi$, and $\varepsilon - 0$ all have mean $O(N^{-1})$. After making the change of variables $y \to \Sigma^{-1}y$, the integral (A.2) is exactly the difference between the two expansions of Theorem 1.

PROOF OF THEOREM 2. Since $E(\text{PCCR}_k) - \rho_k = O(N^{-1})$, it follows that

$$\text{Var}(\text{PCCR}_k) = E\left[\left\{\int_B h^*(y)\, dy - \rho_k\right\}^2\right] + O(N^{-2}). \tag{A.3}$$

It is also easy to see that

$$\text{Var}(\text{ACCR}_k) = (1 - n_k^{-1})E(I_1I_2) + n_k^{-1}E(I_1) - \{E(I_1)\}^2, \tag{A.4}$$

where I_i is the indicator of whether the sample rule classifies X_{ki} correctly. Theorem 1 gives the $O(N^{-1})$ terms of $E(I_1)$. To compute $E(I_1I_2)$, make the transformation (3.1) to (Y_1, Y_2) and obtain the joint density

$$\begin{aligned} m(y_1, y_2) &= h^*(y_1)h^*(y_2)\{1 + \ell n_k^{-1} - (2n)^{-1}\ell(\ell+3) \\ &\quad - (2n_k)^{-1}(B_{11} + 2B_{12} + B_{22}) + (\ell+3)(2n)^{-1}(B_{11} + B_{22}) \\ &\quad - (2n)^{-1}B_{12}^2 - (4n)^{-1}(B_{11}^2 + B_{22}^2)\} + G', \end{aligned} \tag{A.5}$$

where $B_{ij} = y_i'\Sigma^{-1}y_j$ and G' satisfies (2.2).

Since

$$\int_{B\times B} h^*(y_1)h^*(y_2)\,dy_1\,dy_2 = \left\{\int_B h^*(y)\,dy\right\}^2,$$

the expression for $\mathrm{Var}(\mathrm{PCCR}_k)$ will aid in the calculation of the mean of the integral of the leading term of (A.5). For the other terms, h^* can be replaced by ϕ since they each have a factor of $O(N^{-1})$. The remaining integrals are now easily computed to finish the proof.

REFERENCES

EFRON, B. (1979). Bootstrap methods: another look at the jacknife. *Ann. Statist.* **7** 1–26.

HILLS, M. (1966). Allocation rules and their error rates. *J. Roy. Statist. Soc. Ser. B* **28** 1–20.

KHATRI, C. G. (1959). On the mutual independence of certain statistics. *Ann. Math. Statist.* **30** 1258–1262. [Correction (1961) *Ann. Math. Statist.* **32,** 1344].

LACHENBRUCH, P. A. (1975). *Discriminant Analysis.* Hafner Press, New York.

LACHENBRUCH, P. A. and MICKEY, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* **10** 1–11.

MCLACHLAN, G. J. (1973). An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. *Austral. J. Statist.* **15** 210–214.

MCLACHLAN, G. J. (1976). The bias of the apparent error rate in discriminant analysis. *Biometrika* **63** 239–44.

MCLACHLAN, G. J. (1980). The efficiency of Efron's "Bootstrap" approach applied to error rate estimation in discriminant analysis. *J. Statist. Comput. Simulation* **11** 273–280.

RAO, C. R. (1954). A general theory of discrimination when the information about alternative population distributions is based on samples. *Ann. Math. Statist.* **25** 651–670.

SCHERVISH, M. J. (1979). Some results on classifying an observation into one of several multivariate normal populations with equal covariance matrices. Ph.D. thesis, Univ. of Illinois, Univ. Microfilms, Ann Arbor.

SCHERVISH, M. J. (1981). Asymptotic expansions for the means and variances of error rates. *Biometrika* **68** 295–299.

CARNEGIE-MELLON UNIVERSITY
DEPARTMENT OF STATISTICS
SCHENLEY PARK
PITTSBURGH, PA 15213