Ann. Inst. Statist. Math.
Vol. 50, No. 3, 417–431 (1998)

THE COVARIANCE ADJUSTED LOCATION LINEAR DISCRIMINANT FUNCTION FOR CLASSIFYING DATA WITH UNEQUAL DISPERSION MATRICES IN DIFFERENT LOCATIONS

CHI-YING LEUNG

Department of Statistics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China

(Received June 26, 1995; revised May 21, 1997)

Abstract. Classification between two populations dealing with both continuous and binary variables is handled by splitting the problem into different locations. Given the location specified by the values of the binary variables, discrimination is performed using the continuous variables. The location probability model with homoscedastic across location conditional dispersion matrices is adopted for this problem. In this paper, we consider presence of continuous covariates with heterogeneous location conditional dispersion matrices. The continuous covariates have equal location specific mean in both populations. Conditional homoscedasticity fails when strong interaction between the continuous and binary variables is present. A plug-in covariance adjusted rule is constructed and its asymptotic distribution is derived. An asymptotic expansion for the overall error rate is given. The result is extended to include binary covariates.

Key words and phrases: Location linear discriminant function, covariance adjustment, heteroscedastic conditional dispersion matrices, overall expected error rate.

1. Introduction

A vast amount of substantive statistical research centers on prediction. A prediction rule given by a function of explanatory variables affecting a response is to be formulated. For a categorical response, the issue becomes identification of category membership often discussed in discriminant analysis or pattern recognition.

The traditional approach imposes normality on the explanatory variables with homogeneous across category dispersion matrices. Such assumptions oversimplify the problem. In many situations, both mixed discrete and continuous explanatory variables and heterogeneous across category dispersion matrices are encountered. Furthermore, a subset of the variables may not possess discriminatory power for group membership because their true means within a given category are identical in the given populations. These variables called covariates in Cochran and Bliss (1948) can play a substantial role in classification.

In this paper, we consider assigning an object to one of two groups, say Π_1 and Π_2 based on an observed random vector measurement $u' = (x', y')$ where $x' = (x_1, x_2, \ldots, x_b)$ is a row vector of b binary variables and $y' = (y^{(1)'}, y^{(2)'})$, $y^{(1)'} = (y_1, \ldots, y_p)$, $y^{(2)'} = (y_{p+1}, \ldots, y_{p+q})$ is a row vector of $p+q$ continuous variables divided into two parts of p and q variables respectively. The first p continuous variables are the discriminators for simplicity. The last q variables are the covariates in the sense that given the observed values x, $E(Y^{(2)} \mid x, \Pi_1) - E(Y^{(2)} \mid x, \Pi_2)$ is a known function of x. In Section 2, a classification model similar to Krzanowski (1975) is formulated. A plug-in rule is constructed using training data from Π_1 and Π_2. Asymptotic expansions of the distribution of the studentized rule under both Π_1 and Π_2 are derived. In addition, an asymptotic expansion of the overall expected error rate is given. The overall expected error rate offers to answer whether regularization as suggested in Friedman (1989) be used in practice. In Section 5, we consider inclusion of binary covariates. To fix ideas, we first consider the case where the covariates consist of continuous variables only.

2. The classification rule

A systematic approach to classification consisting of mixed binary and continuous variables is based on the location model of Olkin and Tate (1961) adopted for classification in Krzanowski (1975). The b binary variables are expressed by a multinomial variable having $r = 2^b$ locations, each location being represented by an incidence vector $Z' = (Z_1, \ldots, Z_r)$ with observed value $z' = (z_1, \ldots, z_r)$. Each Z_m assumes either the value 0 or 1 and only one nonzero value is allowed in each location. Specifically, let $Z \mid \Pi_i \sim$ Multinomial $(1; p_{1i}, \ldots, p_{ri})$ with probability function $f(z \mid p_i) = \prod_{m=1}^{r} p_{mi}^{z_m}$, $p_i' = (p_{1i}, p_{2i}, \ldots, p_{ri})$,

$$\begin{aligned}
&0 < p_{mi} = E(Z_m \mid \Pi_i) < 1, \quad \sum_{m=1}^{r} z_m = 1, \quad \sum_{m=1}^{r} p_{mi} = 1, \qquad \text{and} \\
&Y \mid \Pi_i, \quad Z_m = 1, \quad Z_k = 0, \\
&m \neq k = 1, \ldots, r \sim N_{p+1}\left(\begin{matrix} p \\ q \end{matrix} \begin{bmatrix} \mu_{mi} \\ \lambda_m \end{bmatrix}, \Sigma^{(m)} \right), \qquad \text{where} \\
&E(Y^{(1)} \mid \Pi_i, Z_m = 1, Z_k = 0, m \neq k = 1, \ldots, r) = \mu_{mi}, \\
&E(Y^{(2)} \mid \Pi_i, Z_m = 1, Z_k = 0, m \neq k = 1, \ldots, r) = \lambda_m,
\end{aligned}$$

λ_m being known. $\Sigma^{(m)}$ is positive definite and partitioned as $\begin{matrix} \\ p \\ q \end{matrix} \begin{matrix} \;\; p \;\;\;\;\; q \\ \begin{bmatrix} \Sigma_{11}^{(m)} & \Sigma_{12}^{(m)} \\ \Sigma_{21}^{(m)} & \Sigma_{22}^{(m)} \end{bmatrix} \end{matrix}$ $k, m = 1, \ldots, r$ and $i = 1, 2$. Given $Z_m = 1$, $m = 1, \ldots, r$, replace Y by $Y - (0', \lambda_m')'$ where $0'$ is a row vector of p zeros, we can assume that λ_m is a column vector of q zeros. When the parameters are known, given $Z_m = 1$, the Bayes rule under specific priors for Π_1 and Π_2 assigns U to Π_1 if and only if $U_m > t$

where $t \in (-\infty, +\infty)$ and

$$U_m = \left[Y^{(1)} - \beta_m Y^{(2)} - \frac{1}{2}(\mu_{m1} + \mu_{m2})\right]' \Sigma_{1.2}^{(m)^{-1}} (\mu_{m1} - \mu_{m2}) - \log(p_{m2}/p_{m1})$$

with $\Sigma_{1.2}^{(m)} = \Sigma_{11}^{(m)} - \Sigma_{12}^{(m)} \Sigma_{22}^{(m)^{-1}} \Sigma_{21}^{(m)}$ and $\beta_m = \Sigma_{12}^{(m)} \Sigma_{22}^{(m)^{-1}}$. The value of t depends on the costs due to misclassification and the priors. In Krzanowski (1975), $\Sigma^{(1)} = \Sigma^{(2)} = \cdots = \Sigma^{(r)}$ is assumed. In the absence of continuous covariates under heterogeneous across location conditional dispersion matrices, Chang and Afifi (1974) derived the rule for $b = 1 (r = 2)$. Balakrishnan *et al.* (1986) obtained the probabilities of misclassification associated with U_m, $m = 1, \ldots, r$ for $p = 1$ and $b = 1$ under the same assumptions. In practice, heterogeneity of across location conditional dispersion matrices often prevails. To implement the procedure, a feasible rule, say $\hat{U}_m$ rather than U_m, $m = 1, \ldots, r$ based on two independent training sets from Π_1 and Π_2 should be constructed. Amongst the existing approaches, the plug-in rule using unbiased estimates for the parameters in U_m, $m = 1, \ldots, r$ is preferred due to its simplicity. Specifically, suppose that large independent samples of size n_1 and n_2 from Π_1 and Π_2 respectively are available with n_{mi} observations from Π_i in location $m, i = 1, 2$; $m = 1, \ldots, r$. Given $Z_m = 1$, $m = 1, \ldots, r$, let

$$\begin{aligned}
&\hat{p}_{mi} = n_{mi}/n_i, \quad i = 1, 2, \\
&Y'_{mji} = (Y_{mji}^{(1)'}, Y_{mji}^{(2)'}), \quad j = 1, \ldots, n_{mi}, \quad i = 1, 2, \\
&\bar{Y}_{mi}^{(h)} = n_{mi}^{-1} \sum_{j=1}^{n_{mi}} Y_{mji}^{(h)}, \quad h = 1, 2, \quad i = 1, 2, \\
&\bar{Y}'_{mi} = (\bar{Y}_{mi}^{(1)'}, \bar{Y}_{mi}^{(2)'}), \quad i = 1, 2.
\end{aligned}$$

Let $\sum_{i=1}^{2} \sum_{j=1}^{n_{mi}} (Y_{mji} - \bar{Y}_{mi})(Y_{mji} - \bar{Y}_{mi})' = \begin{matrix} & p & q \\ p & S_{11}^{(m)} & S_{12}^{(m)} \\ q & S_{21}^{(m)} & S_{22}^{(m)} \end{matrix}$ and $n(m) = n_{m1} + n_{m2} - 2$. Given $Z_m = 1$, unbiased estimates for β_m, $\Sigma^{(m)}$ and $\Sigma_{1.2}^{(m)}$ are given respectively by

$$\begin{aligned}
&\hat{\beta}_m = S_{12}^{(m)} S_{22}^{(m)^{-1}}, \\
&\hat{\Sigma}^{(m)} = (n(m))^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_{mi}} (Y_{mji} - \bar{Y}_{mi})(Y_{mji} - \bar{Y}_{mi})', \quad \text{and} \\
&\hat{\Sigma}_{1.2}^{(m)} = (n(m) - q)^{-1} (S_{11}^{(m)} - S_{12}^{(m)} S_{22}^{(m)^{-1}} S_{21}^{(m)}).
\end{aligned}$$

Unbiased covariates adjusted estimate of μ_{mi} is $\hat{\mu}_{mi} = \bar{Y}_{mi}^{(1)} - \hat{\beta}_m \bar{Y}_{mi}^{(2)}$, $m = 1, \ldots, r$; $i = 1, 2$. The plug-in rule of U_m, say $\hat{U}_m$ is formed by substituting above estimators in U_m for $m = 1, \ldots, r$.

3. Distribution of the studentized classification rule

In this section, we derive an asymptotic expansion of the distribution of $\hat{\Delta}_m^{-1}(\hat{U}_m + (-1)^i \frac{\hat{\Delta}_m^2}{2} + \log(\hat{p}_{m2}/\hat{p}_{m1}))$ under Π_i, with $\Delta_m^2 = (\hat{\mu}_{m1} - \hat{\mu}_{m2})' \hat{\Sigma}_{1.2}^{(m)^{-1}} \cdot (\hat{\mu}_{m1} - \hat{\mu}_{m2})$, $i = 1, 2$; $m = 1, \ldots, r$. The derivation relies on the following lemmas. Their proofs are given in the appendix.

Lemma 1. *Suppose that the following conditions hold:*

Assumption 1. $\frac{n_{m2}}{n_{m1}} \xrightarrow[n_1, n_2 \to \infty]{P} k_m > 0$, $m = 1, \ldots, r$ where $\xrightarrow[n_1, n_2 \to \infty]{P}$ denotes convergence in probability.

Assumption 2. $\frac{n_{s1}}{n_{m1}} \xrightarrow[n_1, n_2 \to \infty]{P} k_{s,m} > 0$, $s, m = 1, \ldots, r$.

Let $n = \sum_{m=1}^r n(m)$. Then

(i) $\frac{n}{n(m)} \xrightarrow[n_1, n_2 \to \infty]{P} 1 + k_m^*$, $k_m^* \geq 0$ $m = 1, \ldots, r$,

(ii) $\lim_{n_1, n_2 \to \infty} \frac{n_2}{n_1} = k > 0$,

(iii) Let $a_m = [(n(m) - q - 1)/(n(m) - 1)]^{1/2}$ and $b_m = q/(n(m) - q - 1)$, then $a_m \xrightarrow[n_1, n_2 \to \infty]{P} 1$ and $b_m \xrightarrow[n_1, n_2 \to \infty]{P} 0$ for $m = 1, \ldots, r$.

Lemma 2. *Let Y be a random vector such that (Y, Z) has a joint probability distribution in each of the r locations specified by the values of $Z' = (Z_1, \ldots, Z_r)$. Suppose that under $Z_m = 1$ the expectation of Y denoted by $E_m(Y)$ exists and is finite for $m = 1, \ldots, r$. Let $E_{2m}(\cdot)$ and $E_{1m}(\cdot)$ denote the operation of taking conditional expectation for given n_{m1} and n_{m2} and taking expectation with respect to the joint distribution of n_{m1} and n_{m2} both under $Z_m = 1$, then $E_m(Y) = E_{1m}(E_{2m}(Y))$, $m = 1, \ldots, r$.*

Lemma 3. *If $\mu_{m1} = 0$, $\mu_{m1} - \mu_{m2} = \delta_m$, $\delta'_m = (\Delta_m, 0, \ldots, 0)$, $0 < \Delta_m = [(\mu_{m1} - \mu_{m2})' \Sigma_{1.2}^{(m)^{-1}} (\mu_{m1} - \mu_{m2})]^{1/2}$ and $\Sigma^{(m)} = I_{p+q}$, a $(p+q) \times (p+q)$ identity matrix, then under the same assumption on (Y, Z) and $Z_m = 1$, for given n_{m1} and n_{m2}, $m = 1, \ldots, r$,*

(i) $E_{2m}(\hat{\mu}_{m1}) = 0$,

(ii) $E_{2m}(\hat{\mu}_{m1} - \hat{\mu}_{m2}) = \delta_m$,

(iii) $E_{2m}(\hat{\beta}_m) = 0$,

(iv) $E_{2m}(\hat{\beta}_m \hat{\beta}'_m) = b_m I_p$,

(v) $E_{2m}(\hat{\Sigma}_{1.2}^{(m)}) = I_p$,

(vi) $E_{2m}(\hat{\mu}_{m1} \hat{\mu}'_{m1}) = n_{m1}^{-1}(1 + b_m) I_p$,

(vii) $E_{2m}((\hat{\mu}_{m1} - \hat{\mu}_{m2} - \delta_m)(\hat{\mu}_{m1} - \hat{\mu}_{m2} - \delta_m)') = (n_{m1}^{-1} + n_{m2}^{-1})(1 + b_m) I_p$,

(viii) $E_{2m}((\hat{\mu}_{m1} - \hat{\mu}_{m2} - \delta_m) \hat{\mu}'_{m1}) = n_{m1}^{-1}(1 + b_m) I_p$,

(ix) $E_{2m}(\delta'_m (\hat{\Sigma}_{1.2}^{(m)} - I_p) \delta_m) = (n(m))^{-1}(p+1) \Delta_m^2$,

(x) $E_{2m}((\delta'_m (\hat{\Sigma}_{1.2}^{(m)} - I_p) \delta_m)^2) = 2(n(m))^{-1} \Delta_m^4$,

(xi) $E_{2m}((\delta'_m(\hat{\beta}_m\hat{\beta}'_m - b_m I_p)\delta_m)^2) = 2d_m\Delta_m^4$, *with* $d_m = q[(n(m)-1)(n(m)-q)^{-1}(n(m)-q-1)^{-1}(n(m)-q-3)^{-1} + (n(m)-2)^{-2}(n(m)-4)^{-1}]$ *where the relevant quantities are defined in Sections* 2 *and* 3.

THEOREM 3.1. *With the same assumptions in Lemma* 1, *for any* $c \in (-\infty, +\infty)$, *given* $Z_m = 1$, *for* $m = 1, \dots, r$ *and* $n = n_1 + n_2 - 2r$,

$$\Pr\left\{\hat{\Delta}_m^{-1}\left(\hat{U}_m - \frac{\hat{\Delta}_m^2}{2} + \log(\hat{p}_{m2}/\hat{p}_{m1})\right) \le c \mid \Pi_1\right\} = \Phi(c) + n^{-1}\phi(c)\psi_{1mc} + O(n^{-2}), \tag{3.1}$$

where $\psi_{1mc} = (1 + k_m^*)[(\frac{p-1}{\Delta_m})(1 + k_m) - (p + \frac{q}{2} - \frac{1}{4} + \frac{k_m}{2})c - \frac{c^3}{4}]$, *and*

$$\Pr\left\{\hat{\Delta}_m^{-1}\left(\hat{U}_m + \frac{\hat{\Delta}_m^2}{2} + \log(\hat{p}_{m2}/\hat{p}_{m1})\right) \le c \mid \Pi_2\right\} = \Phi(c) - n^{-1}\phi(c)\psi_{2mc} + O(n^{-2}), \tag{3.2}$$

where $\psi_{2mc} = (1 + k_m^*)[(\frac{p-1}{\Delta_m})(1 + \frac{1}{k_m}) + (p + \frac{q}{2} - \frac{1}{4} + \frac{1}{2k_m})c + \frac{c^3}{4}]$ *with* $\Phi(\cdot)$ *and* $\phi(\cdot)$ *being the cumulative distribution function and the probability density for* $N(0,1)$ *respectively.*

PROOF OF THEOREM 3.1. Given $Z_m = 1$, $m = 1, \dots, r$ invariance consideration in Memon and Okamoto (1970) is applied to each $\hat{U}_m$ to simplify the distribution of $\hat{U}_m$. Without loss of generality, we assume that $\mu_{m1} = 0$, $\mu_{m2} = -\delta_m$, $\delta'_m = (\Delta_m, 0, \dots, 0)$ and $\Sigma^{(m)} = I_{p+q}$, $m = 1, \dots, r$. Let T_m, W_m, H_m and V_m be defined by

$$\begin{aligned}
\hat{\mu}_{m1} - \hat{\mu}_{m2} &= \delta_m + (n(m))^{-1/2}T_m, \\
\hat{\mu}_{m1} &= (n(m))^{-1/2}W_m, \\
\hat{\beta}_m\hat{\beta}'_m &= b_m I_p + (n(m))^{-1/2}H_m, \\
\hat{\Sigma}_{1.2}^{(m)} &= I_p + (n(m))^{-1/2}V_m.
\end{aligned}$$

By Lemma 3, $E_{2m}(T_m) = 0$, $E_{2m}(W_m) = 0$, $E_{2m}(H_m) = 0$ and $E_{2m}(V_m) = 0$. Given $Z_m = 1$, $m = 1, \dots, r$ and $c \in (-\infty, +\infty)$, by conditional normality of $\hat{U}_m + \log(\hat{p}_{m2}/\hat{p}_{m1})$, and Lemma 2,

$$\Pr\left\{\hat{\Delta}_m^{-1}\left(\hat{U}_m - \frac{\hat{\Delta}_m^2}{2} + \log(\hat{p}_{m2}/\hat{p}_{m1})\right) \le c \mid \Pi_1\right\} = E_m(\Phi(G_m)) = E_{1m}(E_{2m}(\Phi(G_m))), \tag{3.3}$$

where $E_m(\cdot)$ is the expectation with respect to the joint distribution of $\hat{p}_{m1}$, $\hat{p}_{m2}$, $\hat{\mu}_{m1}$, $\hat{\mu}_{m2}$, $\hat{\beta}_m$, and $\hat{\Sigma}_{1.2}^{(m)}$ under Π_1,

$$G_m = \frac{c\hat{\Delta}_m + (\hat{\mu}_{m1} - \hat{\mu}_{m2})'\hat{\Sigma}_{1.2}^{(m)^{-1}}(\hat{\mu}_{m1} - \mu_{m1})}{[(\hat{\mu}_{m1} - \hat{\mu}_{m2})'\hat{\Sigma}_{1.2}^{(m)^{-1}}(I_p + \hat{\beta}_m\hat{\beta}'_m)\hat{\Sigma}_{1.2}^{(m)^{-1}}(\hat{\mu}_{m1} - \hat{\mu}_{m2})]^{1/2}}, \tag{3.4}$$

and $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Expanding $\hat{\Sigma}_{1.2}^{(m)^{-1}}$ and $\hat{\Sigma}_{1.2}^{(m)^{-2}}$ and re-expressing $\hat{\mu}_{m1}$, $\hat{\mu}_{m2}$ and $\hat{\beta}_m$, for fixed large n_{m1} and n_{m2}, we obtain

$$G_m = a_m c + (n(m))^{-1/2} L_m + (n(m))^{-1} Q_m + r_{1m}, \tag{3.5}$$

with

$$L_m = \frac{a_m}{\Delta_m}\left[\delta_m' W_m + \frac{c\delta_m' V_m \delta_m}{2\Delta_m} - \frac{c a_m^2 \delta_m' H_m \delta_m}{2\Delta_m}\right], \quad \text{and} \tag{3.6}$$

$$\begin{aligned} Q_m = \frac{a_m}{\Delta_m}\Bigg[& T_m' W_m - \delta_m' V_m W_m + \frac{c}{2\Delta_m}(\delta_m' V_m^2 \delta_m + T_m' T_m - 2\delta_m' V_m T_m) \\ & - \frac{c}{8\Delta_m^3}(2\delta_m' T_m - \delta_m' V_m \delta_m)^2\Bigg] \\ & + c\Delta_m\Bigg[\frac{-a_m^3}{2\Delta_m^3}\{a_m{}^2(T_m' T_m - 4\delta_m' V_m T_m + 3\delta_m' V_m^2 \delta_m) \\ & \qquad + 2\delta_m' H_m T_m - 3\delta_m'(H_m V_m + V_m H_m)\delta_m\} \\ & \qquad + \frac{3a_m^5}{2\Delta_m^5}\left\{a_m^{-2}(\delta_m' T_m - \delta_m' V_m \delta_m) + \frac{\delta_m' H_m \delta_m}{2}\right\}^2\Bigg] \\ & - \frac{a_m^3}{\Delta_m^3}\left[a_m^{-2}(\delta_m' T_m - \delta_m' V_m \delta_m) + \frac{\delta_m' H_m \delta_m}{2}\right] \\ & \cdot \left[\delta_m' W_m + \frac{c}{2\Delta_m}(2\delta_m' T_m - \delta_m' V_m \delta_m)\right], \end{aligned} \tag{3.7}$$

where r_{1m} is a remainder term such that $E_{1m}(E_{2m}(r_{1m})) = O(n^{-2})$. From Anderson ((1973), p. 968), we have

$$\begin{aligned} & E_{1m}(E_{2m}(\Phi(G_m))) \\ & \quad - E_{1m}(\Phi(a_m c)) \\ & \quad + E_{1m}\Big(\phi(a_m c)\Big\{(n(m))^{-1/2} E_{2m}(L_m) \\ & \qquad + (n(m))^{-1}\left[E_{2m}(Q_m) - \left(\tfrac{a_m c}{2}\right) E_{2m}(L_m^2)\right]\Big\}\Big) \\ & \quad + O(n^{-2}), \end{aligned} \tag{3.8}$$

where $\phi(\cdot)$ denotes the probability density for the standard normal distribution.

Using Lemma 1 and Lemma 3, it can be shown that

$$E_{1m}(\Phi(a_m c)) = \Phi(c) - \frac{\phi(c)}{2} cq(1 + k_n^*) + O(n^{-2}) \quad \text{and} \tag{3.9}$$

$$\begin{aligned} & E_{1m}\Big(\phi(a_m c)\Big\{(n(m))^{-1/2} E_{2m}(L_m) \\ & \qquad + (n(m))^{-1}\left[E_{2m}(Q_m) - \left(\frac{a_m c}{2}\right) E_{2m}(L_m^2)\right]\Big\}\Big) \\ & = n^{-1}\phi(c)(1 + k_m^*)\left[\left(\frac{p-1}{\Delta_m}\right)(1 + k_m) - \left(p + \frac{k_m}{2} - \frac{1}{4}\right)c - \frac{c^3}{4}\right] \\ & \quad + O(n^{-2}). \end{aligned} \tag{3.10}$$

Now (3.1) follows by adding (3.9), (3.10) and a remainder term of order $O(n^{-2})$. Notice that interchanging m_1 and m_2 in $\hat{U}_m$ changes $\hat{U}_m$ to $-\hat{U}_m$. Hence (3.2) follows by subtracting from one the result of substituting $-c$ for c and k_m^{-1} for k_m in the expression after the equality sign in (3.1).

Remark 1. When $q=0$, $b=0(r=1)$, $k_1^*=0$, $k_1=k>0$, $\Delta_1=\Delta>0$ and $p_{11}=p_{12}=1$, (3.1) and (3.2) reduce respectively to (29) and (30) in Anderson ((1973), pp. 969–970).

Remark 2. Results in Memon and Okamoto (1970) and Kanazawa and Fujikoshi (1977) can be derived by setting $b=0(r=1)$, $k_1^*=0$, $k_1=k>0$, $\Delta_1=\Delta>0$ and $p_{11}=p_{12}=1$, in (3.1) and (3.2).

Remark 3. Studentization avoids limitations in Vlachonikolis (1985) when Δ_m is unknown and the threshold t is nonzero.

4 Probabilities of misclassification

In this section, we study the expected error rates. For our plug-in rule, given $Z_m=1$, the probabilities of misclassification between Π_1 and Π_2 in location m are $e_{im}(t)=\Pr\{(-1)^i\hat{U}_m>(-1)^i t \mid \Pi_i\}$, $m=1,\ldots,r$; $i=1,2$ and $t\in(-\infty,+\infty)$ with asymptotic expansions given below.

THEOREM 4.1. *Under the same assumptions stated in Lemma* 1, *given* $Z_m=1$, $t\in(-\infty,+\infty)$, $m=1,\ldots,r$

$$e_{1m}(t)=\Phi(\eta_{1mt})+n^{-1}\phi(\eta_{1mt})(\alpha_{1mt}+\tau_{1mt}+\gamma_{1mt})+O(n^{-2}),\quad \text{and} \tag{4.1}$$

$$e_{2m}(t)=\Phi(\eta_{2mt})+n^{-1}\phi(\eta_{2mt})(\alpha_{2mt}+\tau_{2mt}+\gamma_{2mt})+O(n^{-2}),\quad \text{where} \tag{4.2}$$

$$\eta_{1mt}=\Delta_m^{-1}\left[t+\log(p_{m2}/p_{m1})-\frac{\Delta_m^2}{2}\right],$$

$$\eta_{2mt}=-\Delta_m^{-1}\left[t+\log(p_{m2}/p_{m1})+\frac{\Delta_m^2}{2}\right],$$

$$\alpha_{1mt}=\frac{(1+k)(1-p_{m1})}{\Delta_m p_{m1}}\left\{\frac{3}{4}-\frac{\Delta_m^{-2}}{2}[t+\log(p_{m2}/p_{m1})]\right\}$$
$$-\frac{(1+k^{-1})(1-p_{m2})}{\Delta_m p_{m2}}\left\{\frac{1}{4}+\frac{\Delta_m^{-2}}{2}[t+\log(p_{m2}/p_{m1})]\right\},$$

$$\alpha_{2mt}=\frac{(1+k^{-1})(1-p_{m2})}{\Delta_m p_{m2}}\left\{\frac{3}{4}+\frac{\Delta_m^{-2}}{2}[t+\log(p_{m2}/p_{m1})]\right\}$$
$$-\frac{(1+k)(1-p_{m1})}{\Delta_m p_{m1}}\left\{\frac{1}{4}-\frac{\Delta_m^{-2}}{2}[t+\log(p_{m2}/p_{m1})]\right\},$$

$$\tau_{1mt}=\frac{-q(1+k_m^*)\left[t+\log(p_{m2}/p_{m1})-\dfrac{\Delta_m^2}{2}\right]}{2\Delta_m},$$

$$\tau_{2mt} = \frac{q(1+k_m^*)\left[t+\log(p_{m2}/p_{m1})+\dfrac{\Delta_m^2}{2}\right]}{2\Delta_m},$$

$$\begin{aligned}\gamma_{1mt} = {} & \frac{(p-1)\Delta_m(1+k_m^*)}{4} + \frac{(p-1)}{4\Delta_m}\{3(1+k)p_{m1}^{-1} - (1+k^{-1})p_{m2}^{-1}\} \\ & - [t+\log(p_{m2}/p_{m1})] \\ & \cdot \left\{\frac{3(p-1)(1+k_m^*)}{2\Delta_m} + \frac{(p-3)}{2\Delta_m^3}[(1+k)p_{m1}^{-1} + (1+k^{-1})p_{m2}^{-1}]\right\} \\ & - \frac{\left[t+\log(p_{m2}/p_{m1}) - \dfrac{\Delta_m^2}{2}\right]}{2\Delta_m} \\ & \cdot \left\{\frac{1}{4}[(1+k)p_{m1}^{-1} + (1+k^{-1})p_{m2}^{-1}] \right. \\ & \quad + \Delta_m^{-2}[t+\log(p_{m2}/p_{m1})][(1+k^{-1})p_{m2}^{-1} - (1+k)p_{m1}^{-1}] \\ & \quad + \Delta_m^{-4}[t+\log(p_{m2}/p_{m1})]^2 \\ & \quad \left. \cdot [(1+k)p_{m1}^{-1} + (1+k^{-1})p_{m2}^{-1} + 2(1+k_m^*)\Delta_m^2]\right\},\end{aligned}$$

$$\begin{aligned}\gamma_{2mt} = {} & \frac{(p-1)\Delta_m(1+k_m^*)}{4} + \frac{(p-1)}{4\Delta_m}\{3(1+k^{-1})p_{m2}^{-1} - (1+k)p_{m1}^{-1}\} \\ & + [t+\log(p_{m2}/p_{m1})] \\ & \cdot \left\{\frac{3(p-1)(1+k_m^*)}{2\Delta_m} + \frac{(p-3)}{2\Delta_m^3}[(1+k)p_{m1}^{-1} + (1+k^{-1})p_{m2}^{-1}]\right\} \\ & + \frac{\left[t+\log(p_{m2}/p_{m1}) + \dfrac{\Delta_m^2}{2}\right]}{2\Delta_m} \\ & \cdot \left\{\frac{1}{4}[(1+k)p_{m1}^{-1} + (1+k^{-1})p_{m2}^{-1}] \right. \\ & \quad + \Delta_m^{-2}[t+\log(p_{m2}/p_{m1})][(1+k^{-1})p_{m2}^{-1} - (1+k)p_{m1}^{-1}] \\ & \quad + \Delta_m^{-4}[t+\log(p_{m2}/p_{m1})]^2 \\ & \quad \left. \cdot [(1+k)p_{m1}^{-1} + (1+k^{-1})p_{m2}^{-1} + 2(1+k_m^*)\Delta_m^2]\right\}.\end{aligned}$$

PROOF. Only $e_{1m}(t)$, $m = 1, \ldots, r$ is derived. (4.2) is obtained by substituting $-t$ for t in (4.1) and interchanging the subscripts n_{m1} and n_{m2} and n_1 and n_2 in α_{1mt}, τ_{1mt}, and γ_{1mt}. With δ_m, T_m, W_m, H_m and V_m, $m = 1, \ldots, r$ defined earlier, given $Z_m = 1$, by Lemma 2,

$$e_{1m}(t) = \Pr\{\hat{U}_m < t \mid \Pi_1\} = E_{1m}(E_{2m}(\Phi(G_m^*))), \quad \text{where} \tag{4.3}$$

$$\begin{aligned} G_m^* = {} & a_m\Delta_m^{-1}\left[t+\log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2}\right] \\ & + (n(m))^{-1/2}L_m^* + (n(m))^{-1}Q_m^* + r_m^*, \end{aligned} \tag{4.4}$$

$$
\begin{aligned}
L_m^* &= a_m \Delta_m^{-1} \left[\delta_m' W_m - \delta_m' T_m + \frac{\delta_m' V_m \delta_m}{2} \right] \\
&\quad - a_m^3 \Delta_m^3 \left[t + \log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2} \right] \\
&\quad \cdot \left[a_m^{-2} (\delta_m' T_m - \delta_m' V_m \delta_m) - \frac{\delta_m' H_m \delta_m}{2} \right], \\
Q_m^* &= a_m \Delta_m^{-1} \left[T_m' W_m - \delta_m' V_m W_m + \delta_m' V_m T_m - \frac{T_m' T_m}{2} - \frac{\delta_m' V_m^2 \delta_m}{2} \right] \\
&\quad - \left[t + \log(\hat{p}_{m1}/\hat{p}_{m2}) - \frac{\Delta_m^2}{2} \right] \\
&\quad \cdot \left[\frac{a_m^3 \Delta_m^{-3}}{2} \{ a_m^{-2} (T_m' T_m - 4\delta_m' V_m T_m + 3\delta_m' V_m^2 \delta_m) \right. \\
&\qquad\qquad + 2\delta_m' H_m T_m - 3\delta_m' (H_m V_m + V_m H_m) \delta_m \} \\
&\qquad \left. - \frac{3 a_m^5 \Delta_m^{-5}}{2} \left\{ a_m^{-2} (\delta_m' T_m - \delta_m' V_m \delta_m) + \frac{\delta_m' H_m \delta_m}{2} \right\}^2 \right] \\
&\quad - a_m^3 \Delta_m^{-3} \left[a_m^{-2} (\delta_m' T_m - \delta_m' V_m \delta_m) + \frac{\delta_m' H_m \delta_m}{2} \right] \\
&\quad \cdot \left[\delta_m' W_m - \delta_m' T_m + \frac{\delta_m' V_m \delta_m}{2} \right],
\end{aligned}
$$

and r_m^* is a remainder term similar to r_{1m} in (3.5). Similar argument shows,

$$
\begin{aligned}
(4.5) \quad & E_{1m}(E_{2m}(\Phi(G_m^*))) \\
&= E_{1m} \left(\Phi \left(a_m \Delta_m^{-1} \left[t + \log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2} \right] \right) \right) \\
&\quad + n^{-1} E_{1m} \left(\xi_{1mt} \phi \left(a_m \Delta_m^{-1} \left[t + \log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2} \right] \right) \right) \\
&\quad + O(n^{-2}), \qquad \text{where}
\end{aligned}
$$

$$
\begin{aligned}
\xi_{1mt} &= \frac{n}{n(m)} \left(\frac{p-1}{4} - \frac{3a_m^4 b_m}{8} \right) a_m \Delta_m \\
&\quad + \frac{n}{n(m)} \frac{(p-1) a_m (1+b_m)}{4\Delta_m} \left(\frac{3n(m)}{n_{m1}} - \frac{n(m)}{n_{m2}} \right) \\
&\quad - \frac{n}{n(m)} a_m (1+b_m) [t + \log(\hat{p}_{m2}/\hat{p}_{m1})] \\
&\quad \cdot \left[\frac{3(p-1)}{2(1+b_m)\Delta_m} - \frac{3a_m^4 n(m) d_m}{4(1+b_m)\Delta_m} + \frac{(p-3)}{2\Delta_m^3} \left(\frac{n(m)}{n_{m1}} + \frac{n(m)}{n_{m2}} \right) \right] \\
&\quad - \frac{n}{n(m)} \frac{a_m^3 (1+b_m)}{2\Delta_m} \left[t + \log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2} \right] \\
&\quad \cdot \left[\frac{a_m^4 b_m \Delta_m^2}{8(1+b_m)} + \frac{1}{4} \left(\frac{n(m)}{n_{m1}} + \frac{n(m)}{n_{m2}} \right) \right.
\end{aligned}
$$

$$+ \Delta_m^{-2}\left(\frac{n(m)}{n_{m2}} - \frac{n(m)}{n_{m1}} - \frac{a_m^4 n(m) d_m \Delta_m^2}{2(1+b_m)}\right)[t + \log(\hat{p}_{m2}/\hat{p}_{m1})]$$
$$+ \Delta_m^{-4}\left(\frac{n(m)}{n_{m1}} + \frac{n(m)}{n_{m2}} + \frac{2\Delta_m^2}{1+b_m} + \frac{a_m^4 n(m) d_m \Delta_m^2}{2(1+b_m)}\right)$$
$$\cdot [t + \log(\hat{p}_{m2}/\hat{p}_{m1})]^2\Bigg].$$

It follows from Lemma 2 that,

$$E_{1m}\left(\Phi\left(a_m\Delta_m^{-1}\left[t + \log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2}\right]\right)\right) = \Phi(\eta_{1mt}) + n^{-1}\phi(\eta_{1mt})(\alpha_{1mt} + \tau_{1mt}) + O(n^{-2}), \quad \text{and} \tag{4.6}$$

$$n^{-1}E_{1m}\left(\xi_{1mt}\psi\left(a_m\Delta_m^{-1}\left[t + \log(\hat{p}_{m2}/\hat{p}_{m1}) - \frac{\Delta_m^2}{2}\right]\right)\right) = n^{-1}\phi(\eta_{1mt})\gamma_{1mt} + O(n^{-2}). \tag{4.7}$$

From (4.5), $e_{1m}(t)$ is the sum of (4.6), (4.7) and a reminder term of order $O(n^{-2})$.

COROLLARY. *For the plug-in rule, the overall expected error rate with equal prior probabilities for* Π_1 *and* Π_2 *with threshold at* t, $t \in (-\infty, +\infty)$ *is given by* $\bar{e}(t) = \sum_{i=1}^2 \sum_{m=1}^r p_{mi} e_{im}(t)/2$ *where* $e_{im}(t)$ $i = 1, 2$; $m = 1, \ldots, r$ *is given in Theorem* 4.1.

Remark 4. Both Theorem 3.1 and Theorem 4.1 are valid under conditional normality of $Y^{(1)}$ given $Y^{(2)}$ with mean linear in $Y^{(2)}$ at each of the r locations. Our treatment simplifies the discussion and follows the formulation in Krzanowski (1975).

Remark 5. Anderson ((1984), Corollary 6.6.1, p. 218) can be obtained by letting $q = 0$, $b = 0(r = 1)$, $t = 0$, $k_1^* = 0$, $p_{11} = p_{12} = 1$, $n_{11} = n_{12} = N$, $n = 2N - 2$, $\Delta_1^2 = (\mu_{11} - \mu_{12})'\Sigma_{1.2}^{(1)^{-1}}(\mu_{11} - \mu_{12}) = \Delta^2$ in (4.1) and (4.2). The result in Vlachonikolis (1985) can be derived from Theorem 4.1 upon substituting $q = 0$, $t = 0$. Theorem 4.1 generalizes Leung (1994) to heterogeneous across location conditional dispersion matrices with location specific continuous covariates including the case where $q = 0$.

Remark 6. Based on an empirical study, Vlachonikolis and Marriott (1982) suggest a regularized rule to handle the problem of heterogeneous across location conditional dispersion matrices when continuous covariates are absent. However, the overall expected error rate for their rule has yet to be worked out.

Remark 7. Theorem 3.1 and Theorem 4.1 are valid only when reasonable numbers of data are available at every location for reliable estimation of the means and covariance matrices. When continuous covariates are present, estimates based on a multivariate linear regression model can be used, namely

$$\begin{matrix} p \\ q \end{matrix}\begin{bmatrix} \mu_{mi} \\ 0 \end{bmatrix} = \nu_i + \sum_{j=1}^{b} \alpha_{i,j}^{*} x_j^{(m)} + \sum_{k=2}^{b} \sum_{j=1}^{k-1} \beta_{i,jk}^{*} x_j^{(m)} x_k^{(m)}$$

where $\alpha_{i,j}^{*} \beta_{i,jk}^{*}$ are unknown $(p+q) \times 1$ vector parameters and $x_j^{(m)}$ is the value of the j-th binary variable in cell m of the underlying contingency table for the binary variables in Π_i, $i = 1, 2$; $j, k = 1, \ldots, b$. Pooling the two residual matrices produces an estimate of $\Sigma^{(m)}$ say $\tilde{\Sigma}^{(m)}$ partitioned as $\tilde{\Sigma}^{(m)} = \begin{matrix} \\ p \\ q \end{matrix}\begin{matrix} p \quad\quad q \\ \begin{bmatrix} \tilde{\Sigma}_{11}^{(m)} & \tilde{\Sigma}_{12}^{(m)} \\ \tilde{\Sigma}_{21}^{(m)} & \tilde{\Sigma}_{22}^{(m)} \end{bmatrix} \end{matrix}$ with an estimate of $\begin{matrix} p \\ q \end{matrix}\begin{bmatrix} \mu_{mi} \\ 0 \end{bmatrix}$ given by say, $\tilde{\theta}_{mi}$ partitioned as $\begin{matrix} p \\ q \end{matrix}\begin{bmatrix} \tilde{\theta}_{mi}^{(1)} \\ \tilde{\theta}_{mi}^{(2)} \end{bmatrix}$, $i = 1, 2$. An estimate of μ_{mi} is given by $\tilde{\mu}_{mi} = \tilde{\theta}_{mi}^{(1)} - \tilde{\beta}_m \tilde{\theta}_{mi}^{(2)}$, $\tilde{\beta}_m = \tilde{\Sigma}_{12}^{(m)} \tilde{\Sigma}_{22}^{(m)^{-1}}$ with an estimate of $\Sigma_{1.2}^{(m)}$ given by $\tilde{\Sigma}_{1.2}^{(m)} = n(m)(n(m) - q)^{-1}(\tilde{\Sigma}_{11}^{(m)} - \tilde{\beta}_m \tilde{\Sigma}_{22}^{(m)} \tilde{\beta}_m')$ provided given $Z_m = 1$, $n_{mi} > 1+p+q+b(b+1)/2$, $i = 1, 2$. A corresponding second-order log-linear model is then fitted to the incidence table of counts in each population via iterative scaling as in Haberman (1972) to obtain an estimate say, $\tilde{p}_{mi}$ of p_{mi}, $m = 1, \ldots, r$, $i = 1, 2$ resulting in the plug-in rule, $\tilde{U}_m = [Y^{(1)} - \tilde{\beta}_m Y^{(2)} - \frac{1}{2}(\tilde{\mu}_{m1} + \tilde{\mu}_{m2})]' \tilde{\Sigma}_{1.2}^{(m)^{-1}} (\tilde{\mu}_{m1} - \tilde{\mu}_{m2}) - \log(\tilde{p}_{m2}/\tilde{p}_{m1})$, $m = 1, \ldots, r$. However, the overall expected error rate associated with the rule obtained is difficult to obtain. Alternatively, Tu and Han (1982) suggest an inverse sampling scheme for the construction of $\hat{U}_m$, $m = 1, \ldots, r$ and provide similar asymptotic results.

Remark 8. Theorem 4.1 is essential for studying the overall expected error rate. If interest lies in the overall apparent error rate, the conditional overall error rate given the training data is our concern. The conditional error rate at each location can be obtained by jack-knifing, cross-validation or bootstrapping. See McLachlan ((1992), pp. 339–362). The overall apparent error rate is the weighted sum of the location specific conditional error rates.

5. Generalization

Now suppose that $b - b_1$ $(b_1 < b)$ of the binary variables $X = (X_1, \ldots, X_b)$ are covariates. Each X_i is assumed to take either the value 0 or 1. Following Krzanowski (1975), express X by a multinomial $Z = (Z_1, \ldots, Z_r)$ with $r = 2^b$ locations. $Z_m = 1$ is the incidence for location m and $m = 1 + \sum_{i=1}^{b} x_i 2^{i-1}$ is uniquely determined from the observed $x, m = 1, \ldots, r$. Without loss of generality, let $Z' = (Z^{(1)'}, Z^{(2)'})$, $Z^{(1)'} = (Z_1, \ldots, Z_{r_1})$, $Z^{(2)'} = (Z_{r_1+1}, \ldots, Z_r)$, $(r_1 < r)$ where $Z^{(1)}$ identifies $r_1 = 2^{b_1}$ noncovariates specific locations. $Z^{(2)}$ is associated with the remaining covariates specific locations. Suppose that $Z \mid \Pi_i \sim \text{Multinomial}(1; p_i^{(1)'}, p^{(2)'})$, with $E(Z^{(1)} \mid \Pi_i) = p_i^{(1)} = (p_{1i}, \ldots, p_{r_1 i})'$ and $E(Z^{(2)} \mid \Pi_i) = p^{(2)} = (p_{r_1+1}, \ldots, p_r)'$, being known, $i = 1, 2$. The probability function of Z under Π_i is $f(z \mid p_i^{(1)}, p^{(2)}) = \prod_{m=1}^{r_1} p_{mi}^{z_m} \prod_{m'=r_1+1}^{r} p_{m'}^{z_{m'}}$,

$$\sum_{m=1}^{r_1} z_m + \sum_{m'=r_1+1}^{r} z_{m'} = 1, \qquad \sum_{m=1}^{r_1} p_{mi} + \sum_{m'=r_1+1}^{r} p_{m'} = 1, \qquad \text{and}$$

$Y \mid \Pi_i,\ Z_m = 1,\ Z_k = 0,\ m \neq k = 1, 2, \ldots, r \sim N_{p+q}(\begin{smallmatrix} p \\ q \end{smallmatrix}\left[\begin{smallmatrix}\mu_{mi} \\ \lambda_m\end{smallmatrix}\right], \Sigma^{(m)}),\ i = 1, 2.$ λ_m is assumed zero, $m = 1, \ldots, r$. Given $Z_m = 1$, $m = 1, \ldots, r$, U is allocated to Π_i if and only if $U_m^* > t$ with

$$U_m^* = \begin{cases} U_m, & m = 1, \ldots, r_1 \\ D_m, & m = r_1 + 1, \ldots, r, \end{cases}$$

where $D_m = [Y^{(1)} - \beta_m Y^{(2)} - \frac{1}{2}(\mu_{m1} + \mu_{m2})]' \Sigma_{1.2}^{(m)^{-1}} (\mu_{m1} - \mu_{m2})$. When the parameters are unknown, a plug-in rule say $\hat{U}_m^*$, $m = 1, \ldots, r$ is constructed from full covariates adjusted estimates. The adjustment involves multivariate linear regression and loglinear or multinomial logit fittings. See Aitkin *et al.* ((1989), p. 228) for example. Setting aside $Y' = (Y^{(1)'}, Y^{(2)'})$, given $Y^{(2)}$, simultaneously fit the multinomial logit models, $\log(p_{mi}(Y^{(2)})/p_{1i}(Y^{(2)})) = \beta_{mi}' \tilde{Y}^{(2)}$ with $\tilde{Y}^{(2)'} = (1, Y^{(2)'})$ and $\beta_{mi}' = (\beta_{0mi}, \beta_{1mi}, \ldots, \beta_{qmi})$, $m = 2, \ldots, r_1$ to the training data on $(Y^{(2)}, Z)$ from Π_i, $i = 1, 2$ to obtain $\hat{p}_{1i}(Y^{(2)}) = (1 + \sum_{m=2}^{r_1} \exp(\hat{\beta}_{mi}' \tilde{Y}^{(2)}))^{-1}$ and $\hat{p}_{mi}(Y^{(2)}) = \hat{p}_{1i}(Y^{(2)}) \exp(\hat{\beta}_{mi}' \tilde{Y}^{(2)})$ with fitted $\hat{\beta}_{mi}' = (\hat{\beta}_{0mi}, \hat{\beta}_{1mi}, \ldots, \hat{\beta}_{qmi})$, $m = 2, \ldots, r_1$; $i = 1, 2$. Given (Y, Z) with $Z_m = 1$, the location probability in location m in Π_i is estimated by

$$p_{mi}^* = \hat{p}_{mi}(Y^{(2)}) \left(1 - \sum_{l=r_1+1}^{r} p_l\right) \Big/ \sum_{h=1}^{r_1} \hat{p}_{hi}(Y^{(2)}), \qquad m = 1, \ldots, r_1; \quad i = 1, 2.$$

Plugging in p_{m1}^*, p_{m2}^* with $\tilde{\mu}_{m1}$, $\tilde{\mu}_{m2}$, $\tilde{\beta}_m$ and $\tilde{\Sigma}_{1.2}^{(m)}$ suggested in Remark 7 yields the rule,

$$\hat{U}_m^* = \begin{cases} \tilde{D}_m - \log(p_{m2}^*/p_{m1}^*), & m = 1, \ldots, r_1 \\ \tilde{D}_m, & m = r_1 + 1, \ldots, r, \end{cases}$$

where $\tilde{D}_m = [Y^{(1)} - \tilde{\beta}_m Y^{(2)} - \frac{1}{2}(\tilde{\mu}_{m1} + \tilde{\mu}_{m2})]' \tilde{\Sigma}_{1.2}^{(m)^{-1}} (\tilde{\mu}_{m1} - \tilde{\mu}_{m2})$.

Notice that in the above construction, continuous covariates are used to estimate the location probabilities. Estimates of the parameters of the two multivariate normal distributions are derived from fully adjusted variables. The method relies on the validity of the $2(r_1 - 1)$ logit models. In view of the nature of $Y^{(2)}$, this assumption seems plausible under homogeneity of $\Sigma_{22}^{(1)}, \ldots, \Sigma_{22}^{(r_1)}$ for both populations. However the computing cost incurred should be seriously considered before adopting the method. Apart from this, the adjustment further complicates the distribution problem. Consequently, Theorem 3.1 and Theorem 4.1 are not applicable. Nevertheless, performance of the plug-in rule can be assessed by the weighted sum of the location specific estimated conditional error rates.

Remark 9. Alternatively, smoothed estimates suggested in Remark 7 with covariates adjusted location probability $p_{mi}^{**} = \tilde{p}_{mi}(1 - \sum_{l=r_1+1}^{r} p_l)/\sum_{h=1}^{r_1} \tilde{p}_{hi}$, $i = 1, 2$ can be substituted in U_m^*, $m = 1, \ldots, r$ for the construction of the plug-in rule.

Acknowledgements

The author likes to thank an associate editor and the referee for careful reading of the manuscript. Their criticism greatly improves the presentation of the result.

Appendix

In this section, proofs of the lemmas are given.

PROOF OF LEMMA 1. Recall $n = \sum_{m=1}^{r} n(m) = n_1 + n_2 - 2r$, $n(m) = n_{m1} + n_{m2} - 2$, $m = 1, \ldots, r$. By weak law of large numbers, $\frac{n(m)}{n_1} \xrightarrow[n_1, n_2 \to \infty]{P} p_{m1}(1 + k_m)$ and $\frac{n(m)}{n_2} \xrightarrow[n_1, n_2 \to \infty]{P} p_{m2}(1 + k_m^{-1})$. (i) follows from $\frac{n}{n(m)} = \frac{n_1}{n(m)} + \frac{n_2}{n(m)} - \frac{2r}{n_1} \cdot \frac{n_1}{n(m)} \xrightarrow[n_1, n_2 \to \infty]{P} p_{mi}^{-1}(1 + k_m)^{-1} + p_{m2}^{-1}(1 + k_m^{-1})^{-1} > 0$. Observe that, $\frac{n_2}{n_1} = (\frac{n_{12}}{n_{11}} + \frac{n_{22}}{n_{21}} \cdot \frac{n_{21}}{n_{11}} + \cdots + \frac{n_{r2}}{n_{r1}} \cdot \frac{n_{r1}}{n_{11}})/(1 + \frac{n_{21}}{n_{11}} + \cdots + \frac{n_{r1}}{n_{11}})$. By weak law of large numbers, $\frac{n_2}{n_1} \xrightarrow[n_1, n_2 \to \infty]{P} \frac{k_1 + k_2 k_{2,1} + \cdots + k_r k_{r,1}}{1 + k_{2,1} + \cdots + k_{r,1}} > 0$. (ii) follows from $\lim_{n_1, n_2 \to \infty} n_2/n_1 = E(p\lim_{n_1, n_2 \to \infty} n_2/n_1) = k > 0$ say. (iii) follows from (i) and the definition of a_m and b_m, $m = 1, \ldots, r$.

PROOF OF LEMMA 2. Obvious.

PROOF OF LEMMA 3.

(i) Given n_{m1}, n_{m2} and $Z_m - 1$, $\bar{Y}_{mi} \mid \Pi_i \sim N_{p+q}(\substack{p \\ q}\left[\substack{\mu_{mi} \\ 0}\right], n_{mi}^{-1} I_{p+q})$, $\bar{Y}_{mi}$ and $\hat{\Sigma}^{(m)}$ are independent. From Kshirsagar ((1972), Theorem 2, p. 112), $S_{11}^{(m)} - S_{12}^{(m)} S_{22}^{(m)^{-1}} S_{21}^{(m)} \sim W_p(I_p, n(m) - q)$, $\hat{\beta}_m \mid S_{22}^{(m)} \sim N_{p,q}(0, I_p, S_{22}^{(m)})$ and $S_{22}^{(m)} \sim W_q(I_q, n(m))$. So, $E_{2m}(\hat{\mu}_{m1}) = E_{2m}([I_p, -\hat{\beta}_m]) E_{2m}(\bar{Y}_{m1}) = \mu_{m1}$.

(ii) follows from $E_{2m}(\hat{\mu}_{m1} - \hat{\mu}_{m2}) = E_{2m}([I_p, \hat{\beta}_m]) E_{2m}(\bar{Y}_{m1} - \bar{Y}_{m2}) = \delta_m$.

(iii) follows from the fact that given $S_{22}^{(m)}$, $E_{2m}(\hat{\beta}_m \mid S_{22}^{(m)}) = 0$ and $E_{2m}(\hat{\beta}_m) = E_{2m, S_{22}^{(m)}}(E_{2m}(\hat{\beta}_m \mid S_{22}^{(m)}))$ where $E_{2m, S_{22}^{(m)}}(\cdot)$ denotes the expectation with respect to $S_{22}^{(m)}$ for given n_{m1} and n_{m2} and $Z_m = 1$.

(iv) From Kshirsagar ((1972), Lemma 10, p. 72), if $W \sim W_p(\Sigma, n)$, $n > p + 1$, then $E(W^{-1}) = (n - p - 1)^{-1} \Sigma^{-1}$. Given $S_{22}^{(m)}$, the p rows of $\hat{\beta}_m$ are independent and identically distributed as $N_q(0, S_{22}^{(m)^{-1}})$. Thus

$$\begin{aligned} E_{2m}(\hat{\beta}_m \hat{\beta}'_m) &= E_{2m, S_{22}^{(m)}}(E_{2m}(\hat{\beta}_m \hat{\beta}'_m \mid S_{22}^{(m)})) = E_{2m, S_{22}^{(m)}}(E_{2m}(\mathrm{Tr}(S_{22}^{(m)^{-1}}) I_p)) \\ &= E_{2m, S_{22}^{(m)}}(\mathrm{Tr}(E_{2m}(S_{22}^{(m)^{-1}})) I_p) = b_m I_p. \end{aligned}$$

(v) follows, since $S_{11}^{(m)} - S_{12}^{(m)} S_{22}^{(m)^{-1}} S_{21}^{(m)} \sim W_p(I_p, n(m) - q)$ for given n_{m1} and n_{m2}.

(vi) follows from

$$E_{2m}(\hat{\mu}_{m1}\hat{\mu}'_{m1}) = E_{2m}([I_p, -\hat{\beta}_m]\bar{Y}_{m1}\bar{Y}'_{m1}[I_p, -\hat{\beta}_m]') \\ = E_{2m}(n_{m1}^{-1}[I_p + \hat{\beta}_m\hat{\beta}'_m]).$$

(vii) and (viii) follow similarly.

(ix) follows by applying to each of the r locations, a similar argument to derive (26) in Anderson ((1973), p. 969).

(x) follows by applying to each of the r locations, a similar argument to derive (27) in Anderson ((1973), p. 969).

(xi) follows from the following result.

(A.1) Let $W : (p+q)\times(p+q)$ be a Wishart matrix distributed as $W_{p+q}(I_{p+q}, n)$. Partition W as $W = \begin{smallmatrix} & p & q \\ p & W_{11} & W_{12} \\ q & W_{21} & W_{22}\end{smallmatrix}$ and $\delta' = (\Delta, 0, \ldots, 0) : p \times 1$, $\Delta > 0$, then $\mathrm{Var}((\delta'\hat{\beta}\hat{\beta}'\delta)) = 2q\Delta^4[(n-1)(n-q)^{-1}(n-q-1)^{-1}(n-q-3)^{-1}+(n-2)^{-2}(n-4)^{-1}]$. To obtain (xi), substitute $n(m)$ for n, Δ_m for Δ and $\hat{\beta}_m$ for $\hat{\beta}$ in (A.1). To prove (A.1), note that $\hat{\beta} \mid W_{22} \sim N_{p,q}(0, I_p, W_{22}^{-1})$ and $W_{22} \sim W_q(I_q, n)$. So $(\delta'\hat{\beta})' \mid W_{22} \sim N_q(0, \Delta^2 W_{22}^{-1})$ with $\mathrm{Var}(\delta'\hat{\beta}\hat{\beta}'\delta) = E_1V_2(\delta'\hat{\beta}\hat{\beta}'\delta) + V_1E_2(\delta'\hat{\beta}\hat{\beta}'\delta)$ where $E_2(\cdot)$ denotes the conditional expectation given W_{22}, $V_2(\cdot)$ denotes the conditional variance given W_{22}, $E_1(\cdot)$ denotes the expectation with respect to W_{22} and $V_1(\cdot)$ denotes the variance with respect to W_{22}. By Searle ((1971), Theorem 1, p. 55), $E_2(\delta'\hat{\beta}\hat{\beta}'\delta) = \Delta^2\,\mathrm{Tr}(W_{22}^{-1})$ with

$$V_1E_2(\delta'\hat{\beta}\hat{\beta}'\delta) = \Delta^4\,\mathrm{Var}(\mathrm{Tr}(W_{22}^{-1})) = \Delta^4\,\mathrm{Var}\left(\sum_{i=1}^{q} X_i^{-1}\right) = \frac{2q\Delta^4}{(n-2)^2(n-4)},$$

where $X_1, \ldots, X_q$ are independently identically distributed as χ_n^2. By Searle ((1971), Corollary 1.2, p. 57), $V_2((\delta'\hat{\beta}\hat{\beta}'\delta)) = 2\Delta^4\,\mathrm{Tr}(W_{22}^{-2})$ with $E_1V_2((\delta'\hat{\beta}\hat{\beta}'\delta)) = 2q\Delta^4(n-1)(n-q)^{-1}(n-q-1)^{-1}(n-q-3)^{-1}$ by Srivastava and Khatri ((1979), problem 3.2(iv), p. 97). (A.1) follows.

References

Aitkin, M. A., Anderson, D. A., Francis, B. J. and Hinde, J. P. (1989). *Statistical Modelling in GLIM*, Clarendon, Oxford.

Anderson, T. W. (1973). An asymptotic expansion of the distribution of the studentized classification statistic W, *Ann. Statist.*, **1**, 964–972.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed., Wiley, New York.

Balakrishnan, N., Kocherlakota, S. and Kocherlakota, K. (1986). On the errors of misclassification based on dichotomous and normal variables, *Ann. Inst. Statist. Math.*, **38**, 529–538.

Chang, P. C. and Afifi, A. A. (1974). Classification based on dichotomous and continuous variables, *J. Amer. Statist. Assoc.*, **69**, 336–339.

Cochran, W. G. and Bliss, C. T. (1948). Discriminant functions with covariance, *Ann. Math. Statist.*, **19**, 151–176.

Friedman, J. H. (1989). Regularized discriminant analysis, *J. Amer. Statist. Assoc.*, **84**, 165–175.

Haberman, S. J. (1972). Log-linear fit for contingency tables, *Algorithm AS* 51, *Appl. Statist.*, **21**, 218–225.

Kanazawa, M. and Fujikoshi, Y. (1977). The distribution of the studentized classification statistic W^* in covariate discriminant analysis, *J. Japan Statist. Soc.*, **7**, 81–88.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables, *J. Amer. Statist. Assoc.*, **70**, 782–790.

Kshirsagar, A. M. (1972). *Multivariate Analysis*, Marcel Dekker, New York.

Leung, C. Y. (1994). The location linear discriminant with covariates, *Comm. Statist. Simulation Comput.*, **23**(4), 1027–1046 (Correction: *ibid.* (1995). **24**(4), 1061–1062).

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

Memon, A. Z. and Okamoto, M. (1970). The classification statistic W^* in covariate discriminant analysis, *Ann. Math. Statist.*, **41**, 1491–1499.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables, *Ann. Math. Statist.*, **32**, 448–465 (Correction: *ibid.* (1965). **36**, 343–344).

Searle, S. R. (1971). *Linear Models*, Wiley, New York.

Srivastara, M. S. and Khatri, C. G. (1979). *An Introduction to Multivariate Statistics*, North-Holland, New York.

Tu, C. T. and Han, C. P. (1982). Discriminant analysis based on binary and continuous variables, *J. Amer. Statist. Assoc.*, **77**, 447–454.

Vlachonikolis, I. G. (1985). On the asymptotic distribution of the location linear discriminant function, *J. Roy. Statist. Soc. Ser. B*, **47**, 498–509.

Vlachonikolis, I. G. and Marriott, F. H. C. (1982). Discrimination with mixed binary and continuous data, *Applied Statistics*, **31**, 23–31.