Higher Order Asymptotic Theory for Discriminant Analysis in Exponential Families of Distributions

MASANOBU TANIGUCHI

Osaka University, Toyonaka 560, Japan

This paper deals with the problem of classifying a multivariate observation X into one of two populations Π_1: $p(\mathbf{x}; w^{(1)}) \in S$ and Π_2: $p(\mathbf{x}; w^{(2)}) \in S$, where S is an exponential family of distributions and $w^{(1)}$ and $w^{(2)}$ are unknown parameters. Let $\mathfrak{J}$ be a class of appropriate estimators $(\hat{w}^{(1)}, \hat{w}^{(2)})$ of $(w^{(1)}, w^{(2)})$ based on training samples. Then we develop the higher order asymptotic theory for a class of classification satistics $D = [\hat{W} \mid \hat{W} = \log\{p(X; \hat{w}^{(1)})/p(X; \hat{w}^{(2)})\}, (\hat{w}^{(1)}, \hat{w}^{(2)}) \in \mathfrak{J}]$. The associated probabilities of misclassification of both kinds $M(\hat{w})$ are evaluated up to second order of the reciprocal of the sample sizes. A classification statistic $\hat{W}$ is said to be second order asymptotically best in D if it minimizes $M(\hat{W})$ up to second order. A sufficient condition for $\hat{W}$ to be second order asymptotically best in D is given. Our results are very general and give us a unified view in discriminant analysis. As special results, the Anderson W, the Cochran and Bliss classification statistic, and the quadratic classification statistic are shown to be second order asymptotically best in D in each suitable classification problem. Also, discriminant analysis in a curved exponential family is discussed.

1. INTRODUCTION

First, consider the problem of classifying an observation $\mathbf{X}$ into one of two populations Π_1: $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ and Π_2: $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$ (we call this problem D1). If all the parameters are known, the log likelihood ratio

$$W(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \Sigma) = [\mathbf{X} - (\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})/2]' \Sigma^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})$$

gives the optimal classification rule which minimizes the associated probabilities of misclassification of both kinds. If all the parameters are unknown, and if the training samples $\mathbf{X}_1^{(1)}, ..., \mathbf{X}_{N_1}^{(1)}$ and $\mathbf{X}_1^{(2)}, ..., \mathbf{X}_{N_2}^{(2)}$ drawn from Π_1 and Π_2, respectively, are available Anderson proposed the plug-in version $W_A = W(\hat{\boldsymbol{\mu}}^{(1)}, \hat{\boldsymbol{\mu}}^{(2)}, \hat{\Sigma})$, where $\hat{\boldsymbol{\mu}}^{(1)}$, $\hat{\boldsymbol{\mu}}^{(2)}$, and $\hat{\Sigma}$ are the best unbiased

Received March 2, 1992; revised March 12, 1993.

AMS 1980 subject classifications: primary 62H30, 62H10; secondary 62E20.

Key words and phrases: discriminant analysis, classification statistics, exponential family of distributions, higher order asymptotic theory, curved exponential family, asymptotic expansion.

0047-259X/94 $6.00

estimators of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$ and Σ, respectively. However, in general statistical theory, the best unbiased estimator of $g(\theta)$ is not equal to g (the best unbiased estimator of θ), where $g(\cdot)$ is a known function. Since the associated probabilities of misclassification are involved functions of the parameters the optimality of Anderson's W_A is not obvious in the sense of the best unbiased estimation.

In the asymptotic approach Okamoto (1963) gave an asymptotic expansion of the distribution of W_A up to terms of second order with respect to $(N_1^{-1}, N_2^{-1}, n^{-1})$, where $n = N_1 + N_2 - 2$. Siotani and Wang (1977) added third order terms to Okamoto's expansion and compared it with that of another method. For the problem of classification between Π_1: $N_p((\boldsymbol{\mu}^{(1)\prime}, \boldsymbol{\eta}')', \Sigma)$ and Π_2: $N_p((\boldsymbol{\mu}^{(2)\prime}, \boldsymbol{\eta}')', \Sigma)$ (we call this problem D2), Memon and Okamoto (1970) derived an asymptotic expansion of the distribution of the Cochran and Bliss classification statistic up to the order of $(N_1^{-2}, N_2^{-2}, n^{-2})$. For the problem of classification between Π_1: $N_p(\boldsymbol{\mu}^{(1)}, \Sigma_1)$ and Π_2: $N_p(\boldsymbol{\mu}^{(2)}, \Sigma_2)$ (we call this problem D3), Han (1970), and Wakaki (1990) derived asymptotic expansions for the quadratic classification statistics up to the order of $(N_1^{-1}, N_2^{-1}, n^{-1})$. Regarding asymptotic expansions in various discriminant problems, Siotani (1982) gave an extensive review. However, it seems that there has been no approach based on higher order asymptotic theory.

In this paper we investigate the problem of classifying an observation $\mathbf{X}$ into one of two populations

$$\Pi_1\colon p(\mathbf{x}; w^{(1)}) \in \mathbf{S}$$

and

$$\Pi_2\colon p(\mathbf{x}; w^{(2)}) \in \mathbf{S},$$

where $\mathbf{S}$ is an exponential family of distributions, and $w^{(1)}$ and $w^{(2)}$ are unknown parameters. Let $\mathfrak{J}$ be a class of appropriate estimators $(\hat{w}^{(1)}, \hat{w}^{(2)})$ of $(w^{(1)}, w^{(2)})$ based on training samples. For simplicity we assume that the sample sizes satisfy $N_1 = n$, $N_2 = cn$, where c is a fixed positive constant. Then we define a class of classification statistics by

$$\mathbf{D} = [\hat{W} \mid \hat{W} = \log\{p(\mathbf{X}; \hat{w}^{(1)})/p(\mathbf{X}; \hat{w}^{(2)})\},\ (\hat{w}^{(1)}, \hat{w}^{(2)}) \in \mathfrak{J}].$$

We evaluate the associated probabilities of misclassification of both kinds $M(\hat{W})$ for $\mathbf{D}$ up to $O(n^{-2})$. Then $\hat{W}$ is said to be kth order asymptotically best in $\mathbf{D}$ if $\hat{W}$ minimizes $M(\hat{W})$ up to $O(n^{-k})$ $(k = 1, 2)$. A sufficient condition for $\hat{W}$ to be kth order asymptotically best in $\mathbf{D}$ is given. As special results, it is shown that the Anderson W, the Cochran and Bliss classification statistic, and the quadratic classification statistic are second order asymptotically best in **D** in the problems D1, D2, and D3, respectively. The merits of our approach are as follows:

(1) Since the family **S** of distributions includes many popular families of distributions (e.g., normal, gamma, binomial, etc.), the results are not restricted to the family of normal distributions.

(2) We can deal with the problems D1, D2, and D3 in a unified theory,

Therefore our results give us a very general and unified view for discriminant analysis.

Recently Amari (1985) developed differential geometry of statistical inference for a curved exponential family **M**. Following his theory, we discuss in Section 4 the problem of classification between

$$\Pi_1 : q(\mathbf{x}; u^{(1)}) \in \mathbf{M} \qquad \text{and} \qquad \Pi_2 : q(\mathbf{x}; u^{(2)}) \in \mathbf{M},$$

where $u^{(1)}$ and $u^{(2)}$ are unknown parameters. Then the higher order asymptotic results are elucidated.

2. Higher Order Asymptotic Theory for Discriminant Analysis in an Exponential Family

A family $\mathbf{S} = \{p(\mathbf{x}; \theta)\}$ of distributions is called an exponential family if the density function can be written in the form

$$p(\mathbf{x}; \theta) = \exp\{\theta^i x_i - \psi(\theta)\} \tag{2.1}$$

with respect to some carrier measure $P(\mathbf{x})$, by choosing an adequate parametrization $\theta = (\theta^i)$ and adequate variables $\mathbf{x} = (x_i)$ (e.g., Amari, 1985). Here we adopt the Einstein summation convention. The parameter θ is called the natural parameter of the exponential family. This family **S** includes many popular families of distributions (e.g., normal, gamma, binomial, etc.).

Example 1. Consider the class of multivariate normal distributions $\mathbf{S} = \{N_p(\boldsymbol{\mu}, \Sigma)\}$, where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)'$ and $\Sigma = \{\sigma_{ij}\}$. The probability density function is given by

$$(2\pi)^{-p/2}\{\det \Sigma\}^{-1/2} \exp\{-(y_i - \mu_i)\,\sigma^{ij}(y_j - \mu_j)/2\}, \tag{2.2}$$

where σ^{ij} is the (i, j)th element of Σ^{-1}. If we define the parameter $\theta = (\theta^1, ..., \theta^r)$, $r = p + p(p+1)/2$, by

$$\theta^j = \mu_i \sigma^{ij}, \qquad j = 1, ..., p,$$
$$\theta^{p+1} = -\tfrac{1}{2}\sigma^{11}, ..., \theta^{2p} = -\tfrac{1}{2}\sigma^{pp},$$
$$\theta^{2p+1} = -\sigma^{12},\ \theta^{2p+2} = -\sigma^{13}, ..., \theta^r = -\sigma^{p-1,p},$$

and the new variable $\mathbf{x} = (x_1, ..., x_r)$ by

$$\begin{aligned} &x_1 = y_1, ..., x_p = y_p,\ x_{p+1} = y_1^2, ..., x_{2p} = y_p^2, \\ &x_{2p+1} = y_1 y_2,\ x_{2p+2} = y_1 y_3, ..., x_r = y_{p-1} y_p, \end{aligned} \tag{2.3}$$

then (2.2) can be rewritten as (2.1), where

$$\psi(\theta) = \frac{p}{2} \log 2\pi + \frac{1}{2} \log \det \tilde{\Sigma} + \frac{1}{2} (\theta^1, ..., \theta^p)\, \tilde{\Sigma}^{-1} (\theta^1, ..., \theta^p)',$$

with

$$\tilde{\Sigma} = \begin{pmatrix} -2\theta^{p+1} & -\theta^{2p+1} & \cdots & -\theta^{3p-1} \\ & \cdot & & \cdot \\ \text{sym} & & \cdot & -\theta^r \\ & & & -2\theta^{2p} \end{pmatrix}.$$

In this case the dominating measure $P(\mathbf{x})$ is concentrated on the manifold defined by the relation (2.3). In what follows, it is convenient to introduce a new coordinate system $w = (w^\alpha) = (w^1, ..., w^r)$, so that $\theta = \theta(w)$. By this coordinate system, we also denote $\mathbf{S} = \{p(\mathbf{x}; w)\}$.

Now consider the problem of classifying a p-dimensional observation into one of two populations

$$\Pi_1\colon p(\mathbf{x}; w^{(1)}) \in \mathbf{S}$$

and

$$\Pi_2\colon p(\mathbf{x}; w^{(2)}) \in \mathbf{S},$$

where $w = (u', v')'$ with $u = (u_1, ..., u_s)'$ and $v = (v_1, ..., v_t)'$, $t + s = r$, reparametrizing by $\theta^i = \theta^i(w)$ for $i = 1, ..., s$ and $\theta^i = \theta^i(v)$ for $i = s+1, ..., r$, and furthermore $w^{(k)} = (u^{(k)\prime}, v')'$, $u^{(k)} = (u_1^{(k)}, ..., u_t^{(k)})'$ $(k = 1, 2)$. This setting includes many classification problems as special cases.

EXAMPLE 2. Classification between $\Pi_1\colon N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ and $\Pi_2\colon N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$, where $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, ..., \mu_p^{(i)})'$ $(i = 1, 2)$ and $\Sigma = \{\sigma_{ij}\}$, becomes our problem if we set $u^{(i)} = (\mu_1^{(i)}, ..., \mu_p^{(i)})'$ with $s = p$ and $v = (\sigma_{11}, ..., \sigma_{pp}, \sigma_{12}, \sigma_{13}, ..., \sigma_{p-1,p})'$ with $t = p(p+1)/2$.

EXAMPLE 3. Classification between Π_1: $N_p((\boldsymbol{\mu}^{(1)'}, \boldsymbol{\eta}')', \Sigma)$ and Π_2: $N_p((\boldsymbol{\mu}^{(2)'}, \boldsymbol{\eta}')', \Sigma)$, where $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, ..., \mu_q^{(i)})'$ $(i = 1, 2)$, $\boldsymbol{\eta} = (\eta_1, ..., \eta_{p-q})'$, and $\Sigma = \{\sigma_{ij}\}$, becomes our problem if we set $u^{(i)} = (\mu_1^{(i)}, ..., \mu_q^{(i)})'$ with $s = q$ and $v = (\eta_1, ..., \eta_{p-q}, \sigma_{11}, ..., \sigma_{pp}, \sigma_{12}, \sigma_{13}, ..., \sigma_{p-1,p})'$ with $t = p - q + p(p+1)/2$.

EXAMPLE 4. Classification between Π_1: $N_p(\boldsymbol{\mu}^{(1)}, \Sigma_1)$ and Π_2: $N_p(\boldsymbol{\mu}^{(2)}, \Sigma_2)$, where $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, ..., \mu_p^{(i)})'$ and $\Sigma_i = \{\sigma_{jl}^{(i)}\}$ $(i = 1, 2)$, also becomes our problem if we set

$$u^{(i)} = (\mu_1^{(i)}, ..., \mu_p^{(i)}, \sigma_{11}^{(i)}, ..., \sigma_{pp}^{(i)}, \sigma_{12}^{(i)}, \sigma_{13}^{(i)}, ..., \sigma_{p-1,p}^{(i)})'$$

with $s = p + p(p+1)/2$ and $t = 0$.

Now we return to the general problem. It is desired that the observation $\mathbf{X}$ be classified into Π_1: $p(\mathbf{x}; w^{(1)}) \in \mathrm{S}$ or Π_2: $p(\mathbf{x}; w^{(2)}) \in \mathrm{S}$. If both the parameters $w^{(1)}$ and $w^{(2)}$ are known the optimal classification rule is based on the statistic

$$W(\mathbf{X}; w^{(1)}, w^{(2)}) = \log\{p(\mathbf{X}; w^{(1)})/p(\mathbf{X}; w^{(2)})\}. \tag{2.4}$$

That is, the classification regions $R_1 = \{\mathbf{x};\ W(\mathbf{x}; w^{(1)}, w^{(2)}) \geqslant 0\}$ and $R_2 = \{\mathbf{x};\ W(\mathbf{x}; w^{(1)}, w^{(2)}) < 0\}$ minimize the misclassification probability

$$\int_{R_2} p(\mathbf{x}; w^{(1)})\, dP(\mathbf{x}) + \int_{R_1} p(\mathbf{x}; w^{(2)})\, dP(\mathbf{x})$$

(e.g., Anderson, 1984). In most applications the parameters $w^{(1)}$ and $w^{(2)}$ are unknown and must be inferred from samples, one from each population. To treat this case we investigate the behavior of the misclassification probability

$$\begin{aligned} &P\{W(\mathbf{X}; \tilde{w}^{(1)}, \tilde{w}^{(2)}) < 0 \mid \Pi_1\} + P\{W(\mathbf{X}; \tilde{w}^{(1)}, \tilde{w}^{(2)}) \geqslant 0 \mid \Pi_2\} \\ &\quad = P^{(2\mid 1)}(\tilde{\lambda} \mid w^{(1)}) + P^{(1\mid 2)}(\tilde{\lambda} \mid w^{(2)}), \qquad \text{say,} \end{aligned}$$

where $\tilde{\lambda} = (\tilde{u}^{(1)'}, \tilde{u}^{(2)'}, \tilde{v}')'$ is a nonrandom point in a neighborhood of $\lambda = (u^{(1)'}, u^{(2)'}, v')'$.

Initially, we make the following assumption.

Assumption 1. The misclassification probability

$$P^{(2\mid 1)}(\tilde{\lambda} \mid w^{(1)}) + P^{(1\mid 2)}(\tilde{\lambda} \mid w^{(2)}) \qquad (= M(\tilde{\lambda}), \text{ say})$$

is five times continuously differentiable with respect to $\tilde{\lambda}$ in a neighborhood of λ.

Let $\tilde{\lambda}^{\alpha}$ and λ^{α} be the αth component of $\tilde{\lambda}$ and λ, respectively, and put $d = \dim \tilde{\lambda} = \dim \lambda$. We abbreviate $\partial/\partial\lambda^{\alpha}$ as ∂_{α} ($\alpha = 1, ..., d$). Since $W(\mathbf{X}; w^{(1)}, w^{(2)})$ gives the best classification region, $M(\tilde{\lambda})$ is minimized at λ. It follows from Assumption 1 that

$$\partial_{\alpha} M(\lambda) = 0, \qquad \alpha = 1, ..., d, \tag{2.5}$$

and

$$\text{the matrix } \{\partial_{\alpha}\partial_{\beta} M(\lambda); \alpha, \beta = 1, ..., d\} \geqslant 0, \tag{2.6}$$

where the inequality $\geqslant 0$ means nonnegative definiteness. By a Taylor expansion we have

$$\begin{aligned} M(\tilde{\lambda}) = {} & M(\lambda) + \tfrac{1}{2}(\tilde{\lambda}^{\alpha} - \lambda^{\alpha})(\tilde{\lambda}^{\beta} - \lambda^{\beta})\{\partial_{\alpha}\partial_{\beta} M(\lambda)\} \\ & + \tfrac{1}{6}(\tilde{\lambda}^{\alpha} - \lambda^{\alpha})(\tilde{\lambda}^{\beta} - \lambda^{\beta})(\tilde{\lambda}^{\gamma} - \lambda^{\gamma})\{\partial_{\alpha}\partial_{\beta}\partial_{\gamma} M(\lambda)\} \\ & + \tfrac{1}{24}(\tilde{\lambda}^{\alpha} - \lambda^{\alpha})(\tilde{\lambda}^{\beta} - \lambda^{\beta})(\tilde{\lambda}^{\gamma} - \lambda^{\gamma})(\tilde{\lambda}^{\delta} - \lambda^{\delta})\{\partial_{\alpha}\partial_{\beta}\partial_{\gamma}\partial_{\delta} M(\lambda)\} \\ & + \text{higher order terms of } (\tilde{\lambda} - \lambda). \end{aligned} \tag{2.7}$$

We treat the case in which we have a sample from each of two exponential family distributions. That is, a sample $\mathbf{X}_1^{(1)}, ..., \mathbf{X}_{N_1}^{(1)}$ is drawn from Π_1: $p(\mathbf{x}; w^{(1)}) \in \mathrm{S}$, and $\mathbf{X}_1^{(2)}, ..., \mathbf{X}_{N_2}^{(2)}$ is drawn from Π_2: $p(\mathbf{x}; w^{(2)}) \in \mathrm{S}$. Let $X_{ij}^{(k)}$ be the jth component of $\mathbf{X}_i^{(k)}$, $k = 1, 2$. Then the log likelihood based on $\mathbf{X}_1^{(1)}, ..., \mathbf{X}_{N_1}^{(1)}$ and $\mathbf{X}_1^{(2)}, ..., \mathbf{X}_{N_2}^{(2)}$ is

$$\begin{aligned} L(u^{(1)}, u^{(2)}, v) = {} & \sum_{j=1}^{s} \theta^{j}(u^{(1)}, v) \sum_{i=1}^{N_1} X_{ij}^{(1)} + \sum_{j=1}^{s} \theta^{j}(u^{(2)}, v) \sum_{i=1}^{N_2} X_{ij}^{(2)} \\ & + \sum_{j=s+1}^{r} \theta^{j}(v) \left\{ \sum_{i=1}^{N_1} X_{ij}^{(1)} + \sum_{i=1}^{N_2} X_{ij}^{(2)} \right\} \\ & - N_1 \psi(\theta(w^{(1)})) - N_2 \psi(\theta(w^{(2)})). \end{aligned} \tag{2.8}$$

To develop the asymptotic theory of estimation of λ based on the training samples, we need the following assumptions.

Assumption 2. $N_1 = n$, $N_2 = cn$, where c is a fixed positive constant.

We deal with a class of estimators $\hat{\lambda} = (\hat{u}^{(1)'}, \hat{u}^{(2)'}, \hat{v}')'$ of λ, which satisfy

Assumption 3.

$$E(\hat{\lambda}) - \lambda = o(1), \tag{2.9}$$

and $\partial_{\alpha} = \partial/\partial\lambda^{\alpha}$ and E are interchangeable.

To avoid many sophisticated regularity conditions we simply assume the validity of the Edgeworth expansion for $\hat{\lambda}$ in the following form (see Bhattacharya and Rao, 1976; Bhattacharya and Ghosh, 1978).

Assumption 4. Let P_n be the probability distribution of $\sqrt{n}(\hat{\lambda}-\lambda)$ $(=v$, say). Then it holds that for every bounded, Borel-measurable function f of v on $\mathbf{R}^d$,

$$\left|\int fd(P_n-F_n)\right|=o(n^{-2}), \tag{2.10}$$

where $F_n=F_n(v)$, $v=(v^1, ..., v^d)'$, is the formal Edgeworth expansion of $\sqrt{n}(\hat{\lambda}-\lambda)$.

Let $\mathfrak{J}$ be the class of estimators $\hat{\lambda}$ satisfying all the assumptions above. We propose the plug-in version of $W(\mathbf{X}; w^{(1)}, w^{(2)})$ as a classification statistic and denote the class of them by

$$\mathbf{D}=\{\hat{W} \mid \hat{W}=W(\mathbf{X}; \hat{w}^{(1)}, \hat{w}^{(2)}), \text{ with } \hat{w}^{(i)}=(\hat{u}^{(i)\prime}, \hat{v}')', \ i=1, 2\}.$$

Sometimes the notation $\hat{W}=\hat{W}(\hat{\lambda})$ is used. Then the misclassification probability based on $\hat{W}\in\mathbf{D}$ is $E\{M(\hat{\lambda})\}$, where the expectation is taken according to the distribution of $\hat{\lambda}$. The purpose of this paper is to determine which $\hat{W}(\hat{\lambda})$ $(\in\mathbf{D})$ minimizes $E\{M(\hat{\lambda})\}$ in the sense of asymptotic theory with respect to n. The following lemma gives a useful evaluation.

Lemma 1. *For* $\hat{\lambda}\in\mathfrak{J}$,

$$\begin{aligned}E\{M(\hat{\lambda})\}&=M(\lambda)+\tfrac{1}{2}\{\partial_\alpha\partial_\beta M(\lambda)\}\, E\{(\hat{\lambda}^\alpha-\lambda^\alpha)(\hat{\lambda}^\beta-\lambda^\beta)\}\\&\quad+\tfrac{1}{6}\{\partial_\alpha\partial_\beta\partial_\gamma M(\lambda)\}\, E\{(\hat{\lambda}^\alpha-\lambda^\alpha)(\hat{\lambda}^\beta-\lambda^\beta)(\hat{\lambda}^\gamma-\lambda^\gamma)\}\\&\quad+\tfrac{1}{24}\{\partial_\alpha\partial_\beta\partial_\gamma\partial_\delta M(\lambda)\}\\&\quad\times E\{(\hat{\lambda}^\alpha-\lambda^\alpha)(\hat{\lambda}^\beta-\lambda^\beta)(\hat{\lambda}^\gamma-\lambda^\gamma)(\hat{\lambda}^\delta-\lambda^\delta)\}+o(n^{-2}),\end{aligned}$$

where $\hat{\lambda}^\alpha$ is the αth element of $\hat{\lambda}$.

We give the proofs of lemmas and theorems in Section 5. Now we can find a classification statistic $\hat{W}(\hat{\lambda})\in\mathbf{D}$ which minimizes $EM(\hat{\lambda})$ up to order $O(n^{-1})$. Such a classification statistic $\hat{W}(\hat{\lambda})$ is said to be first order asymptotically best. Discussion in the proof of Lemma 1 leads to

$$\begin{aligned}E\{M(\hat{\lambda})\}&=M(\lambda)+\frac{1}{2n}\{\partial_\alpha\partial_\beta M(\lambda)\}\, E\{\sqrt{n}(\hat{\lambda}^\alpha-\lambda^\alpha)\sqrt{n}(\hat{\lambda}^\beta-\lambda^\beta)\}\\&\quad+o(n^{-1}), \qquad \text{for } \hat{\lambda}\in\mathfrak{J}.\end{aligned} \tag{2.11}$$

In view of (2.11), $\hat{W}(\hat{\lambda})$ is first order asymptotically best if $\hat{\lambda}$ minimizes

$$\{\partial_\alpha \partial_\beta M(\lambda)\}\ E\{\sqrt{n}(\hat{\lambda}^\alpha - \lambda^\alpha)\sqrt{n}(\hat{\lambda}^\beta - \lambda^\beta)\}. \tag{2.12}$$

Let $\mathscr{F}$ be a $(d \times d)$-matrix whose (α, β)th element is $I_{\alpha\beta} = n^{-1}E(\partial_\alpha L)(\partial_\beta L)$, where L is the abbreviated notation of (2.8) and $\partial_\alpha = \partial/\partial\lambda^\alpha$. Then we have

THEOREM 1. *A classification statistic $\hat{W}(\hat{\lambda})$ is first order asymptotically best in* **D** *if the estimator $\hat{\lambda}$ has the stochastic expansion*

$$\sqrt{n}(\hat{\lambda}^\alpha - \lambda^\alpha) = I^{\alpha\beta}\frac{1}{\sqrt{n}}\partial_\beta L + o_p(1), \qquad \alpha = 1, ..., d, \tag{2.13}$$

where $I^{\alpha\beta}$ is the (α, β)th element of $\mathscr{F}^{-1}$.

The following corollary holds since the first order stochastic expansion of the maximum likelihood estimator is given by (2.13) (e.g., Takeuchi and Morimune, 1985, p. 190).

COROLLARY 1. *Let $\hat{\lambda}_{\mathrm{ML}}$ be the maximum likelihood estimator of λ. Then $\hat{W}(\hat{\lambda}_{\mathrm{ML}})$ is first order asymptotically best in* **D**.

We proceed to discuss the higher order asymptotic optimality in a class of classification statistics. Let $\mathscr{Y}$ be the class of estimators $\hat{\lambda}$ which satisfy the following conditions:

(i) $\hat{\lambda}$ has the stochastic expansion

$$\sqrt{n}(\hat{\lambda}^\alpha - \lambda^\alpha) = Z^\alpha + \frac{1}{\sqrt{n}}Q^\alpha + o_p(n^{-1/2}), \qquad \alpha = 1, ..., d, \tag{2.14}$$

where $Z^\alpha = I^{\alpha\beta}n^{-1/2}\,\partial L/\partial\lambda^\beta$, and $Q^\alpha = O_p(1)$.

(ii) The second order bias of $\hat{\lambda}$ is evaluated as

$$E\{\sqrt{n}(\hat{\lambda}^\alpha - \lambda^\alpha)\} = \mu^\alpha(\lambda)/\sqrt{n} + o(n^{-1}), \qquad \alpha = 1, .., d, \tag{2.15}$$

where $\mu^\alpha(\lambda) = E(Q^\alpha\}$, and is continuously differentiable with respect to λ.

We also introduce a class of bias-adjusted estimators defined by

$$\mathscr{Y}^* = \left\{\hat{\lambda}^* = (\hat{\lambda}^{1*}, ..., \hat{\lambda}^{d*})' \,\middle|\, \hat{\lambda}^{\alpha*} = \hat{\lambda}^\alpha - \frac{1}{n}\mu^\alpha(\hat{\lambda}),\ \hat{\lambda} \in \mathscr{Y}\right\}.$$

For this class $\mathscr{Y}^*$, define the class of classification statistics by

$$\mathscr{F}^* = \{\hat{W}^* \mid \hat{W}^* = \hat{W}(\hat{\lambda}^*),\ \hat{\lambda}^* \in \mathscr{Y}^*\}.$$

If $\hat{W}(\hat{\lambda}^*) \in \mathscr{F}^*$ minimizes $E\{M(\hat{\lambda})\}$ up to order $O(n^{-2})$, it is said to be second order asymptotically best in $\mathscr{F}^*$.

THEOREM 2. (1) *For $\hat{\lambda}^* \in \mathscr{Y}^*$, it holds that*

$$\begin{aligned} E\{M(\hat{\lambda}^*)\} = M(\lambda) &+ \frac{1}{2n}\{\partial_\alpha \partial_\beta M(\lambda)\}\cdot I^{\alpha\beta} \\ &+ \frac{1}{2n^2}\{\partial_\alpha \partial_\beta M(\lambda)\}\{\mathrm{Cov}(Q^\alpha, Q^\beta)\} + \frac{1}{6n^2}\{\partial_\alpha \partial_\beta \partial_\gamma M(\lambda)\}\cdot c^{\alpha\beta\gamma} \\ &+ \frac{1}{24n^2}\{\partial_\alpha \partial_\beta \partial_\gamma \partial_\delta M(\lambda)\}\{I^{\alpha\beta}I^{\gamma\delta} + I^{\alpha\gamma}I^{\beta\delta} + I^{\alpha\delta}I^{\beta\gamma}\} + o(n^{-2}), \end{aligned} \tag{2.16}$$

where $c^{\alpha\beta\gamma}$ are independent of the choice of an estimator in $\mathscr{Y}^*$.

(2) *Let $\hat{\lambda}^*_{\mathrm{ML}}$ be the bias-adjusted maximum likelihood estimator of λ. Then $\hat{W}^* = \hat{W}(\hat{\lambda}^*_{\mathrm{ML}})$ is second order asymptotically best in $\mathscr{F}^*$.*

It may be noted that (2.16) describes the higher order asymptotic structure of $E\{M(\hat{\lambda}^*)\}$ clearly. Concrete examples of our results are given in the next section.

3. EXAMPLES

In this section we show that the results in Section 2 give a unified view of many multivariate works done for Examples 2, 3, and 4 and that some famous classification statistics are second order asymptotically best in $\mathscr{F}^*$. For a general review on asymptotic expansions of classification statistics we may refer to Siotani (1982).

(I) *Classification between Π_1: $N_p(\boldsymbol{\mu}^{(1)}, \Sigma)$ and Π_2: $N_p(\boldsymbol{\mu}^{(2)}, \Sigma)$ stated in Example* 2. Let

$$W_{\mathrm{A}} = [\mathbf{X} - \tfrac{1}{2}(\bar{\mathbf{X}}^{(1)} + \bar{\mathbf{X}}^{(2)})]'\, S^{-1}(\bar{\mathbf{X}}^{(1)} - \bar{\mathbf{X}}^{(2)}),$$

where

$$\bar{\mathbf{X}}^{(i)} = N_i^{-1} \sum_{j=1}^{N_i} \mathbf{X}_j^{(i)} \qquad (i = 1, 2),$$

$$\begin{aligned} S = \frac{1}{N_1 + N_2 - 2}\Bigg\{&\sum_{j=1}^{N_1} (\mathbf{X}_j^{(1)} - \bar{\mathbf{X}}^{(1)})(\mathbf{X}_j^{(1)} - \bar{\mathbf{X}}^{(1)})' \\ &+ \sum_{j=1}^{N_2} (\mathbf{X}_j^{(2)} - \bar{\mathbf{X}}^{(2)})(\mathbf{X}_j^{(2)} - \bar{\mathbf{X}}^{(2)})'\Bigg\}, \end{aligned} \tag{3.1}$$

which is known to be Anderson's classification statistic (e.g., Anderson, 1984. Okamoto (1963) gave the second order asymptotic expansion of W_A and further, Siotani and Wang (1977) gave the third order one. In our notation we can set

$$\lambda = (\boldsymbol{\mu}^{(1)\prime}, \boldsymbol{\mu}^{(2)\prime}, (\operatorname{vec} \Sigma)')', \qquad \tilde{\lambda} = (\tilde{\boldsymbol{\mu}}^{(1)\prime}, \tilde{\boldsymbol{\mu}}^{(2)\prime}, (\operatorname{vec} \tilde{\Sigma})')',$$

where $\operatorname{vec} \Sigma = (\sigma_{11}, ..., \sigma_{pp}, \sigma_{12}, \sigma_{13}, ..., \sigma_{p-1,p})'$ and $\operatorname{vec} \tilde{\Sigma} = (\tilde{\sigma}_{11}, ..., \tilde{\sigma}_{pp}, \tilde{\sigma}_{12}, \tilde{\sigma}_{13}, ..., \tilde{\sigma}_{p-1,p})'$. Then,

$$M(\tilde{\lambda}) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi b}} \exp\{-(x-a_1)^2/2b\}\, dx + \int_{0}^{\infty} \frac{1}{\sqrt{2\pi b}} \exp\{-(x-a_2)^2/2b\}\, dx,$$

where

$$a_i = [\boldsymbol{\mu}^{(i)} - \tfrac{1}{2}(\tilde{\boldsymbol{\mu}}^{(1)} + \tilde{\boldsymbol{\mu}}^{(2)})]' \tilde{\Sigma}^{-1}(\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)}) \qquad (i = 1, 2),$$
$$b = (\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)})' \tilde{\Sigma}^{-1} \Sigma \tilde{\Sigma}^{-1}(\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)}).$$

Also, the maximum likelihood estimators of $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, and Σ are given by

$$\hat{\boldsymbol{\mu}}^{(1)}_{\mathrm{ML}} = \bar{\mathbf{X}}^{(1)}, \qquad \hat{\boldsymbol{\mu}}^{(2)}_{\mathrm{ML}} = \bar{\mathbf{X}}^{(2)},$$

and

$$\hat{\Sigma}_{\mathrm{ML}} = \frac{N_1 + N_2 - 2}{N_1 + N_2} S,$$

respectively. It is easy to see that a bias-adjusted maximum likelihood estimator of Σ is given by

$$\hat{\Sigma}^{*}_{\mathrm{ML}} = \hat{\Sigma}_{\mathrm{ML}} + \frac{2}{N_1 + N_2} \hat{\Sigma}_{\mathrm{ML}},$$

and

$$S - \hat{\Sigma}^{*}_{\mathrm{ML}} = O_p(n^{-2}). \tag{3.2}$$

In view of (3.2), the difference between S and $\hat{\Sigma}^{*}_{\mathrm{ML}}$ does not disturb our asymptotic theory up to $O(n^{-2})$. Therefore, from Theorem 2 Anderson's W_A is shown to be second order asymptotically best in $\mathscr{F}^{*}$.

Next we discuss the following covariate discriminant problem.

(II) *Classification of* $\mathbf{X} = (x', y')'$ *into one of the two populations* $\Pi_1 : N_p((\boldsymbol{\mu}^{(1)'}, \boldsymbol{\eta}')', \Sigma)$ *and* $\Pi_2 : N_p((\boldsymbol{\mu}^{(2)'}, \boldsymbol{\mu}')', \Sigma)$ *stated in Example* 3. Let $\mathbf{X}_1^{(1)} = (\mathbf{x}_1^{(1)'}, \mathbf{y}_1^{(1)'})', \ldots, \mathbf{X}_{N_1}^{(1)} = (\mathbf{x}_{N_1}^{(1)'}, \mathbf{y}_{N_1}^{(1)'})'$ and $\mathbf{X}_1^{(2)} = (\mathbf{x}_1^{(2)'}, \mathbf{y}_1^{(2)'})', \ldots, \mathbf{X}_{N_2}^{(2)} = (\mathbf{x}_{N_2}^{(2)'}, \mathbf{y}_{N_2}^{(2)'})'$ be the training samples from Π_1 and Π_2, respectively. Cochran and Bliss (1946) proposed the classification statistic

$$W_{\mathrm{CB}}^* = [\mathbf{x}^* - \tfrac{1}{2}(\bar{\mathbf{x}}_1^* + \bar{\mathbf{x}}_2^*)]' \, S_{11.2}^{-1}(\bar{\mathbf{x}}_1^* - \bar{\mathbf{x}}_2^*),$$

where $\mathbf{x}^* = \mathbf{x} - S_{12}S_{22}^{-1}y$, $\bar{\mathbf{x}}_i^* = \bar{\mathbf{x}}_i - S_{12}S_{22}^{-1}\bar{\mathbf{y}}_i$, $\bar{\mathbf{x}}_i = \sum_{k=1}^{N_i} \mathbf{x}_k^{(i)}/N_i$, $\bar{\mathbf{y}}_i = \sum_{k=1}^{N_i} \mathbf{y}_k^{(i)}/N_i$, $S_{11.2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$, and

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

which is given by the same definition as (3.1). Memon and Okamoto (1970) gave the second order asymptotic expansion of the distribution of W_{CB}^*. In this case

$$W(\mathbf{X}; w^{(1)}, w^{(2)}) = [\mathbf{x} - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \boldsymbol{\eta}) - \tfrac{1}{2}(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)})]' \, \Sigma_{11.2}^{-1}(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}), \tag{3.3}$$

where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

We set

$$\lambda = (\boldsymbol{\mu}^{(1)'}, \boldsymbol{\mu}^{(2)'}, \boldsymbol{\eta}', (\operatorname{vec}\Sigma)')', \qquad \tilde{\lambda} = (\tilde{\boldsymbol{\mu}}^{(1)'}, \tilde{\boldsymbol{\mu}}^{(2)'}, \tilde{\boldsymbol{\eta}}', (\operatorname{vec}\tilde{\Sigma})')',$$

where

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{pmatrix}.$$

Then,

$$M(\tilde{\lambda}) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi h}} \exp\{-(x - c_1)^2/2h\}\, dx + \int_{0}^{\infty} \frac{1}{\sqrt{2\pi h}} \exp\{-(x - c_2)^2/2h\}\, dx,$$

where

$$c_i = [\boldsymbol{\mu}^{(i)} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}(\boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}) - \tfrac{1}{2}(\tilde{\boldsymbol{\mu}}^{(1)} + \tilde{\boldsymbol{\mu}}^{(2)})]' \, \tilde{\Sigma}_{11.2}^{-1}(\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)}) \quad (i = 1, 2),$$

$$h = (\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)})' \, \tilde{\Sigma}_{11.2}^{-1}(I, -\tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}) \, \Sigma (I, -\tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1})' \, \tilde{\Sigma}_{11.2}^{-1}(\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)}),$$

$\tilde{\Sigma}_{11.2} = \tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21}$, and I is the $q \times q$ identity matrix. The likelihood of the training samples is

$$\begin{aligned} L = {} & (2\pi)^{-p(N_1+N_2)/2} \, [|\Sigma_{11.2}| \cdot |\Sigma_{22}|]^{-(N_1+N_2)/2} \exp\Bigg[-\frac{1}{2} \operatorname{tr} \Sigma_{11.2}^{-1} \\ & \times \Bigg\{ \sum_{i=1}^{2} \sum_{k=1}^{N_i} [\mathbf{x}_k^{(i)} - \boldsymbol{\mu}^{(i)} - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_k^{(i)} - \boldsymbol{\eta})] \\ & \times [\mathbf{x}_k^{(i)} - \boldsymbol{\mu}^{(i)} - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_k^{(i)} - \boldsymbol{\eta})]' \Bigg\} \\ & -\frac{1}{2} \operatorname{tr} \Sigma_{22}^{-1} \Bigg\{ \sum_{i=1}^{2} \sum_{k=1}^{N_i} (\mathbf{y}_k^{(i)} - \boldsymbol{\eta})(\mathbf{y}_k^{(i)} - \boldsymbol{\eta})' \Bigg\} \Bigg]. \end{aligned} \tag{3.4}$$

By a slight modification of Fujikoshi and Kanazawa (1976) we can see that the maximization of L with respect to $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\boldsymbol{\eta}$, $\Sigma_{12}\Sigma_{22}^{-1}$, and $\Sigma_{11.2}^{-1}$ is achieved at

$$\begin{aligned} \hat{\boldsymbol{\mu}}^{(1)} &= \bar{\mathbf{x}}_1 - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}(\hat{\mathbf{y}}_1 - \hat{\boldsymbol{\eta}}), \\ \hat{\boldsymbol{\mu}}^{(2)} &= \bar{\mathbf{x}}_2 - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}(\bar{\mathbf{y}}_2 - \hat{\boldsymbol{\eta}}), \\ \hat{\boldsymbol{\eta}} &= (N_1 + N_2)^{-1} \left(\sum_{k=1}^{N_1} \mathbf{y}_k^{(1)} + \sum_{k=1}^{N_2} \mathbf{y}_k^{(2)} \right), \end{aligned}$$

$$\widehat{\Sigma_{12}\Sigma_{22}^{-1}} = \widehat{\Omega}_{12}\hat{\Omega}_{22}^{-1}, \qquad \widehat{\Sigma_{11.2}^{-1}} = (\hat{\Omega}_{11} - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{\Omega}_{21})^{-1},$$

where

$$\hat{\Omega} = \begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{pmatrix} = \frac{N_1 + N_2 - 2}{N_1 + N_2} S.$$

As in the previous example it is shown that the Cochran and Bliss classification statistic W_{CB}^* is asymptotically equivalent to the plug-in version $\hat{W}(\tilde{\lambda}_{\mathrm{ML}}^*)$ of W. Thus it follows from Theorem 2 that W_{CB}^* is second order asymptotically best in $\mathscr{F}^*$.

We end this section with the following example.

(III) *Classification between Π_1: $N_p(\boldsymbol{\mu}^{(1)}, \Sigma_1)$ and Π_2: $N_p(\boldsymbol{\mu}^{(2)}, \Sigma_2)$ stated in Example* 4. In this case,

$$\begin{aligned} W(\mathbf{X}; w^{(1)}, w^{(2)}) = {} & -\tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}^{(1)})' \Sigma_1^{-1}(\mathbf{X} - \boldsymbol{\mu}^{(1)}) + \tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}^{(2)})' \Sigma_2^{-1}(\mathbf{X} - \boldsymbol{\mu}^{(2)}) \\ & - \tfrac{1}{2} \log |\Sigma_1 \Sigma_2^{-1}| = -\tfrac{1}{2} Q(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)}, \Sigma_1, \Sigma_2) \qquad \text{(say)}. \end{aligned}$$

When Σ_1 and Σ_2 are known and proportional, Han (1969) derived an asymptotic expansion of the distribution of $Q(\bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}, \Sigma_1, \Sigma_2)$ up to $O(n^{-2})$. For the case in which Σ_1 and Σ_2 are unknown circular matrices, Han (1970) gave an asymptotic expansion of the distribution of $Q(\bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}, S_1, S_2)$ up to $O(n^{-1})$, where

$$S_i = \frac{1}{N_i - 1} \sum_{j=1}^{N_i} (\mathbf{X}_j^{(i)} - \bar{\mathbf{X}}^{(i)})(\mathbf{X}_j^{(i)} - \bar{\mathbf{X}}^{(i)})' \qquad (i = 1, 2).$$

Under the assumption that Σ_1 and Σ_2 are proportional and unknown, Wakaki (1990) derived an asymptytotic expansion for $Q(\bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}, S_1, S_2)$ up to $O(n^{-1})$. Returning to our general result in Theorem 2 we can see that the plug-in version $\hat{W}(\hat{\lambda}^*_{\mathrm{ML}})$ of W is asymptotically equivalent to $-\frac{1}{2}Q(\bar{\mathbf{X}}^{(1)}, \bar{\mathbf{X}}^{(2)}, S_1, S_2)$ up to $O(n^{-2})$ without the circular or the proportional assumption. Therefore all Q's in the works cited above are second order asymptotically best in $\mathscr{F}^*$. However, if we do not assume some special covariance structure, the explicit expression for $M(\lambda)$ is usually difficult to obtain. If we assume $\Sigma_i = \gamma_i^{-1} I_p$, $\gamma_i > 0$ $(i = 1, 2)$, where I_p is the $p \times p$ identity matrix (see Wakaki, 1990), we can evaluate $M(\lambda)$ as follows. Let

$$\lambda = (\boldsymbol{\mu}^{(1)\prime}, \boldsymbol{\mu}^{(2)\prime}, \gamma_1, \gamma_2)',$$

$$\tilde{\lambda} = (\tilde{\boldsymbol{\mu}}^{(1)\prime}, \tilde{\boldsymbol{\mu}}^{(2)\prime}, \tilde{\gamma}_1, \tilde{\gamma}_2)',$$

$$\tilde{a} = \frac{1}{2}(\tilde{\gamma}_2 - \tilde{\gamma}_1),$$

$$\tilde{\delta} = \gamma_i \left(\boldsymbol{\mu}^{(i)} - \frac{\tilde{\gamma}_1 \tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\gamma}_2 \tilde{\boldsymbol{\mu}}^{(2)}}{\tilde{\gamma}_1 - \tilde{\gamma}_2}\right)' \left(\boldsymbol{\mu}^{(i)} - \frac{\tilde{\gamma}_1 \tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\gamma}_2 \tilde{\boldsymbol{\mu}}^{(2)}}{\tilde{\gamma}_1 - \tilde{\gamma}_2}\right),$$

$$\tilde{b}_i = \frac{\gamma_i \tilde{\gamma}_1 \tilde{\gamma}_2}{2(\tilde{\gamma}_1 - \tilde{\gamma}_2)} (\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)})' (\tilde{\boldsymbol{\mu}}^{(1)} - \tilde{\boldsymbol{\mu}}^{(2)}) - \frac{p\gamma_i}{2} \log \tilde{\gamma}_1/\tilde{\gamma}_2$$
$$- \frac{\gamma_i^2 \gamma_2}{2} (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})' (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}) + \frac{p\gamma_i}{2} \log \gamma_1/\gamma_2.$$

Then it is not difficult to show that

$$M(\tilde{\lambda}) = \int_{-\infty}^{0} \varphi_1(x; \tilde{\lambda})\, dx + \int_{0}^{\infty} \varphi_2(x; \tilde{\lambda})\, dx,$$

where $\varphi_i(x; \tilde{\lambda})$ is the probability density function having the characteristic function

$$e^{it\tilde{b}_i}(1 - 2it\tilde{a})^{-p/2} \exp\{it\tilde{\delta}_i \tilde{a}/(1 - 2it\tilde{a})\}, \qquad t \in \mathbf{R}^1.$$

4. Discriminant Analysis in a Curved Exponential Family

Amari (1985) developed differential geometry of statistical inference for a curved exponential family

$$\mathbf{M} = [q(\mathbf{x}; u) \mid q(\mathbf{x}; u) = \exp\{\theta^i(u)\, x_i - \psi(\theta(u))\}]$$

embedded in the exponential family

$$\mathbf{S} = [p(\mathbf{x}; \theta) \mid p(\mathbf{x}; \theta) = \exp\{\theta^i x_i - \psi(\theta)\}],$$

where $u = (u^1, ..., u^q)'$ and $q < r = \dim \theta$. Here the term "curved" stems from the requirement that the dimension of the minimal sufficient statistic exceed that of the parameter space. In this section we consider the problem of classifying an observation $\mathbf{X}$ into one of two populations Π_1: $q(\mathbf{x}; u^{(1)}) \in \mathbf{M}$ and Π_2: $q(\mathbf{x}; u^{(2)}) \in \mathbf{M}$. where $u^{(i)} = (u^{(i)1}, ..., u^{(i)q})'$ $(i = 1, 2)$ are unknown, and $u^{(1)} \neq u^{(2)}$. Suppose that $\mathbf{X}_1^{(1)}, ..., \mathbf{X}_{N_1}^{(1)}$ and $\mathbf{X}_1^{(2)}, ..., \mathbf{X}_{N_2}^{(2)}$ are the training samples drawn from Π_1 and Π_2, respectively. Henceforth we make the following assumptions;

(i) $N_1 = cN_2$ for some constant $c > 0$.

(ii) The partial derivatives $\partial/\partial\theta^i$, $\partial/\partial u^i$, etc., exist up to necessary orders and are interchangeable with E.

(iii) The Edgeworth expansions for the estimators $\tilde{\theta}$, $\hat{u}$, etc., are valid.

The log likelihood based on the training samples is written as

$$\begin{aligned} L(u^{(1)}, u^{(2)}) = {} & \sum_{j=1}^{r} \theta^j(u^{(1)}) \sum_{i=1}^{N_1} X_{ij}^{(1)} + \sum_{j=1}^{r} \theta^j(u^{(2)}) \sum_{i=1}^{N_2} X_{ij}^{(2)} \\ & - N_1 \psi(\theta(u^{(1)})) - N_2 \psi(\theta(u^{(2)})), \end{aligned} \tag{4.1}$$

where $X_{ij}^{(k)}$ is the jth element of $\mathbf{X}_i^{(k)}$ $(k = 1, 2)$. Sufficient statistics here are

$$T_j^{(k)} = \frac{1}{N_k} \sum_{i=1}^{N_k} X_{ij}^{(k)}, \qquad k = 1, 2\ j = 1, ..., r.$$

Define $T^{(k)} = (T_1^{(k)}, ..., T_r^{(k)})'$ $(k = 1, 2)$. It is easy to show that

$$E(T_j^{(1)}) = \partial_j \psi(\theta(u^{(1)})) = \eta_j^{(1)} \qquad \text{(say)},$$
$$E(T_j^{(2)}) = \partial_j \psi(\theta(u^{(2)})) = \eta_j^{(2)} \qquad \text{(say)},$$

where $\partial_j = \partial/\partial\theta^j$ (see Amari, 1985). The parameters $\eta^{(k)} = (\eta_1^{(k)}, ..., \eta_r^{(k)})'$ $(k = 1, 2)$ are called the expectation parameters. Since they depend on the unknown parameters $u^{(k)}$ $(k = 1, 2)$, we sometimes use the notations $\eta^{(k)} = \eta^{(k)}(u^{(k)})$ $(k = 1, 2)$.

Turning to the estimation problem, consider the estimation of $u^{(k)}$ ($k=1,2$) in $\mathbf{M}$ by $T^{(k)}$ ($k=1,2$) in $\mathbf{S}$. This leads us to the equations

$$T^{(1)}=\eta^{(1)}(\hat{u}^{(1)}), \qquad T^{(2)}=\eta^{(2)}(\hat{u}^{(2)}),$$

which cannot be solved in $\hat{u}^{(k)}$ ($k=1,2$), since $\dim T^{(k)}=r>q=\dim u^{(k)}$. Therefore we introduce new extra coordinates $v^{(k)}=(v^{(k),1},\dots,v^{(k),r-q})'$ ($k=1,2$) so that $w^{(k)}=(w^{(k),1},\dots,w^{(k),r})'=(u^{(k)'},v^{(k)'})'$ become coordinate systems in $\mathbf{S}$. Then the equations

$$T^{(1)}=\eta^{(1)}(\hat{u}^{(1)},\hat{v}^{(1)}), \qquad T^{(2)}=\eta^{(2)}(\hat{u}^{(2)},\hat{v}^{(2)}) \tag{4.2}$$

can be uniquely solved with respect to $\hat{u}^{(k)}$ and $\hat{v}^{(k)}$, where we note that the conditions $\eta^{(k)}(u^{(k)},0)=\eta^{(k)}(u^{(k)})$ ($k=1,2$) are assumed. Fixing $u^{(1)}$ and $u^{(2)}$, we locally define

$$A(u^{(1)},u^{(2)})=\{\boldsymbol{\eta}\mid\boldsymbol{\eta}=(\eta^{(1)'},\eta^{(2)'})',\ u^{(1)}=\hat{u}^{(1)}(\eta^{(1)}),\ u^{(2)}=\hat{u}^{(2)}(\eta^{(2)})\},$$

which is called the ancillary family associated with the estimators $\hat{u}^{(1)}$ and $\hat{u}^{(2)}$. We can see that the determination of the estimators $\hat{u}^{(k)}$ of $u^{(k)}$ is in one to one correspondence with the set $A(u^{(1)},u^{(2)})$.

For our classification problem,

$$W_{\mathrm{c}}(\mathbf{X};u^{(1)},u^{(2)})=\log\{q(\mathbf{X};u^{(1)})/q(\mathbf{X};u^{(2)})\}$$

gives the optimal classification rule if $u^{(1)}$ and $u^{(2)}$ are known. Define

$$M_{\mathrm{c}}(\tilde{\lambda}_{\mathrm{c}})=P\{W_{\mathrm{c}}(\mathbf{X};\tilde{u}^{(1)},\tilde{u}^{(2)})<0\mid\Pi_1\}+P\{W_{\mathrm{c}}(\mathbf{X};\tilde{u}^{(1)},\tilde{u}^{(2)})\geqslant 0\mid\Pi_2\},$$

where $\tilde{\lambda}_{\mathrm{c}}=(\tilde{u}^{(1)'},\tilde{u}^{(2)'})'$ is a nonrandom point in a neighbourhood of $\lambda_{\mathrm{c}}=(u^{(1)'},u^{(2)'})'$. We set

$$\begin{aligned}\mathscr{S}_{\mathrm{c}}=\{\hat{\lambda}_{\mathrm{c}}\mid\hat{\lambda}_{\mathrm{c}}=(\hat{u}^{(1)'},\hat{u}^{(2)'})',\ &\hat{u}^{(1)}\text{ and }\hat{u}^{(2)}\text{ are defined by (4.2)}\\ &\text{and }\boldsymbol{\eta}\in A(u^{(1)},u^{(2)})\text{ if }\eta^{(k)}(u^{(k)})\in\mathbf{M}\ (k=1,2)\},\\ \mathscr{D}_{\mathrm{c}}=\{\hat{W}_{\mathrm{c}}\mid\hat{W}_{\mathrm{c}}=W_{\mathrm{c}}(\mathbf{X};\hat{u}^{(1)},\hat{u}^{(2)}),\ &\hat{\lambda}_{\mathrm{c}}=(\hat{u}^{(1)'},\hat{u}^{(2)'})'\in\mathscr{S}_{\mathrm{c}}\}.\end{aligned}$$

Sometimes the notation $\hat{W}_{\mathrm{c}}=\hat{W}_{\mathrm{c}}(\hat{\lambda}_{\mathrm{c}})$ is used. The classification statistic $\hat{W}_{\mathrm{c}}(\hat{\lambda}_{\mathrm{c}})$ is called kth order asymptotically best if it minimizes $EM_{\mathrm{c}}(\hat{\lambda}_{\mathrm{c}})$ up to order $O(n^{-k})$ ($k=1,2$). We state the following theorem without a proof because the discussion of Section 3, together with that of Chapter 5 in Amari (1985), proves it with a slight modification.

THEOREM 3. *Assume that*

$$g^{(k)}_{a\kappa}=E\left\{\frac{\partial}{\partial u^{(k),a}}\log q(\mathbf{X};\eta^{(k)})\,\frac{\partial}{\partial v^{(k),\kappa}}\log q(\mathbf{X};\eta^{(k)})\right\}=0, \tag{4.3}$$

for $k=1,2$; $a=1,\ldots,q$, *and* $\kappa=1,\ldots,r-q$, *where* $g_{a\kappa}^{(k)}$ *are evaluated at* $\eta^{(k)}=\eta^{(k)}(u^{(k)},0)$. (*We simply denote this assumption by* $A(u^{(1)},u^{(2)})\perp\mathbf{M}$.)

(1) *If an estimator* $\hat\lambda_c\in\mathscr{S}_c$ *satisfies* $A(u^{(1)},u^{(2)})\perp\mathbf{M}$, *then* $\hat W_c=\hat W_c(\hat\lambda_c)$ *is first order asymptotically best in* $\mathscr{D}_c$. *In particular, if* $\hat\lambda_{c\mathrm{ML}}$ *is the maximum likelihood etimator of* λ_c, *then* $\hat W_c(\hat\lambda_{c\mathrm{ML}})$ *is first order asymptotically best in* $\mathscr{D}_c$.

(2) *Let* $\mathscr{Y}_c^*$ *be the class of estimators* $\hat\lambda_c^*$ *which satisfy* $A(u^{(1)},u^{(2)})\perp\mathbf{M}$ *and are bias-adjusted in the manner of Section* 2, *and define*

$$\mathscr{F}_c^*=\{\hat W_c^*\mid \hat W_c^*==\hat W_c(\hat\lambda_c^*),\ \hat\lambda_c^*\in\mathscr{Y}_c^*\}.$$

If $\hat\lambda_{c\mathrm{ML}}^*$ *is the bias-adjusted maximum likelihood estimator of* λ_c, *then* $\hat W_c(\hat\lambda_{c\mathrm{ML}}^*)$ *is second order asymptotically best in* $\mathscr{F}_c^*$.

EXAMPLE 5. Consider the problem of classification between Π_1: $N_p(\boldsymbol{\mu}(u^{(1)}),\Sigma(u^{(1)}))$ and Π_2: $N_p(\boldsymbol{\mu}(u^{(2)}),\Sigma(u^{(2)}))$, where the functional forms of $\boldsymbol{\mu}(\cdot)$ and $\Sigma(\cdot)$ are known, and $\dim u^{(1)}=\dim u^{(2)}<p+p(p+1)/2$. It is assumed that $u^{(1)}\neq u^{(2)}$ and $u^{(1)}$ and $u^{(2)}$ are unknown. The model $N_p(\boldsymbol{\mu}(u^{(1)}),\Sigma(u^{(1)}))$ includes the factor analysis model, etc., as special cases (see Lawley and Maxwell, 1971). All the results in Theorem 3 can be applied to this problem.

5. PROOFS

Proof of Lemma 1. Since $M(\hat\lambda)$ is the sum of two probabilities, it is bounded by 2 almost surely. It follows from Assumption 4 that

$$E\{M(\hat\lambda)\}=\int_{\mathbf{R}^d}M\left(\lambda+\frac{1}{\sqrt{n}}v\right)dF_n(v)+o(n^{-2}). \tag{5.1}$$

Using a fundamental property of the Edgeworth approximation, if we choose a positive constant c_0 appropriately, we obtain

$$\int_{A_n^c}M\left(\lambda+\frac{1}{\sqrt{n}}v\right)dF_n(v)\leqslant 2\int_{A_n^c}dF_n(v)=o(n^{-2}). \tag{5.2}$$

where $A_n=\{v\mid |v^1|\leqslant c_0\sqrt{\log n},\ldots,|v^d|\leqslant c_0\sqrt{\log n}\}$. Hence, by (5.1)

$$E\{M(\hat\lambda)\}=\int_{A_n}M\left(\lambda+\frac{1}{\sqrt{n}}v\right)dF_n(v)+o(n^{-2}). \tag{5.3}$$

We expand $M(\lambda+(1/\sqrt{n})v)$ of (5.3) in the fifth order Taylor series, leading to

$$E\{M(\hat{\lambda})\}=\int_{A_n}[M(\lambda)+v^{\alpha}\{\partial_{\alpha}M(\lambda)\}/\sqrt{n}+v^{\alpha}v^{\beta}\{\partial_{\alpha}\partial_{\beta}M(\lambda)\}/2n$$

$$+\cdots+\frac{1}{4!n^2}v^{\alpha}v^{\beta}v^{\gamma}v^{\delta}\{\partial_{\alpha}\partial_{\beta}\partial_{\gamma}\partial_{\delta}M(\lambda)\}]\,dF_n(v)+o(n^{-2}). \quad (5.4)$$

Recall that $\partial_{\alpha}M(\lambda)=0$, and note Lemma 2.1 of Bhattacharya and Ghosh (1978) to complete the proof.

Proof of Theorem 1. Differentiation of (2.9) with respect to λ^{α} yields

$$E\left\{\sqrt{n}(\hat{\lambda}^{\beta}-\lambda^{\beta})\frac{1}{\sqrt{n}}(\partial_{\alpha}L)\right\}=\delta(\alpha,\beta)+o(1), \quad (5.5)$$

where $\delta(\alpha,\beta)$ is Kronecker's delta. Let M and $V(\hat{\lambda})$ be the $(d\times d)$-matrices $\{\partial_{\alpha}\partial_{\beta}M(\lambda)\}$ and $\{E\{\sqrt{n}(\hat{\lambda}^{\alpha}-\lambda^{\alpha})\sqrt{n}(\hat{\lambda}^{\beta}-\lambda^{\beta})\}\}$, respectively. From (2.12), if $\hat{\lambda}$ minimizes $\operatorname{tr}MV(\hat{\lambda})$ subject to (5.5), then $\hat{W}(\hat{\lambda})$ gives the first order asymptotically best classification. Since $M\geqslant 0$ by (2.6), a solution of our problem is given by the solution of the problem

$$\begin{array}{c}\min_{\hat{\lambda}}V(\hat{\lambda})\\ \text{subject to (5.5),}\end{array} \quad (5.6)$$

where min is taken in the sense of the nonnegative definiteness $\geqslant$. From (5.5), the solution must satisfy

$$\sqrt{n}(\hat{\lambda}^{\alpha}-\lambda^{\alpha})=A^{\alpha\beta}\frac{1}{\sqrt{n}}(\partial_{\beta}L)+\text{lower order terms} \quad (5.7)$$

for some matrix $\{A^{\alpha\beta}\}$. Substituting (5.7) into (5.5) we have $A^{\alpha\beta}=I^{\alpha\beta}$.

Proof of Theorem 2. We do not give the details of calculation because they are very troublesome, and the methods are essentially similar to those of Takeuchi and Morimune (1985, p. 192), Amari (1985), and Taniguchi (1991). Let $S^{\alpha}=\sqrt{n}(\hat{\lambda}^{\alpha *}-\lambda^{\alpha})$, $\alpha=1,...,d$. Then it can be shown that

$$E(S^{\alpha}S^{\beta})=I^{\alpha\beta}+\frac{1}{n}\operatorname{Cov}(Q^{\alpha},Q^{\beta})+o(n^{-1}), \quad (5.8)$$

$$\begin{aligned}E(S^{\alpha}S^{\beta}S^{\gamma})&=\operatorname{cum}(S^{\alpha},S^{\beta},S^{\gamma})+o(n^{-1/2})\\&=\frac{1}{2\sqrt{n}}[I^{\alpha\xi}\partial_{\xi}I^{\beta\gamma}+I^{\beta\xi}\partial_{\xi}I^{\alpha\gamma}+I^{\gamma\xi}\partial_{\xi}I^{\alpha\beta}]\\&\quad-\frac{1}{2}E\{Z^{\alpha}Z^{\beta}Z^{\gamma}\}+o(n^{-1/2}),\end{aligned} \quad (5.9)$$

where cum{ , , } is the joint cumulant of { , , }. Here $E\{Z^\alpha Z^\beta Z^\gamma\} = O(n^{-1/2})$ and is independent of the choice of an estimator. Also, we can show that

$$E(S^\alpha S^\beta S^\gamma S^\delta) = \text{cum}(S^\alpha, S^\beta)\,\text{cum}(S^\gamma, S^\delta) + \text{cum}(S^\alpha, S^\gamma)\,\text{cum}(S^\beta, S^\delta) + \text{cum}(S^\alpha, S^\delta)\,\text{cum}(S^\beta, S^\gamma) + O(n^{-1}). \tag{5.10}$$

Combining (5.8), (5.9), and (5.10) with Lemma 1, we obtain (2.16) for $\hat{\lambda}^* \in \mathscr{Y}^*$.

Turning to the proof of (2), let

$$Z_\alpha = n^{-1/2}\partial_\alpha L, \qquad Z_{\alpha\beta} = n^{-1/2}\{\partial_\alpha\partial_\beta L - E\partial_\alpha\partial_\beta L\},$$

$$R_{\alpha\beta\gamma} = E\left(\frac{1}{n}\,\partial_\alpha\partial_\beta\partial_\gamma L\right), \qquad J_{\alpha\beta\gamma} = E(Z_\alpha Z_{\beta\gamma}),$$

$$K_{\alpha\beta\gamma} = \sqrt{n}E(Z_\alpha Z_\beta Z_\gamma), \qquad M_{\alpha\beta\cdot\gamma\delta} = E(Z_{\alpha\beta} Z_{\gamma\delta}).$$

Then, for $\hat{\lambda}^* \in \mathscr{Y}^*$, the following relations hold:

$$E(Z_\beta Z_\gamma \tilde{Q}^\alpha) = -I^{\alpha\alpha'}(J_{\alpha'\beta\gamma} + K_{\alpha'\beta\gamma}) + o(1), \tag{5.11}$$

$$E(Z_\beta Z_{\gamma\delta} \tilde{Q}^\alpha) = -I^{\alpha\eta}\partial_\beta I_{\eta\gamma'} I^{\gamma'\alpha'} J_{\alpha'\gamma\delta} + I^{\alpha\alpha'} M_{\gamma\delta\cdot\alpha'\beta} + o(1). \tag{5.12}$$

Here $\tilde{Q}^\alpha = Q^\alpha - \mu^\alpha(\hat{\lambda})$. In Eq. (2.16), only the matrix $H(Q) = \{\text{Cov}(Q^\alpha, Q^\beta)\}$ depends on the choice of an estimator. Therefore we are led to the problem

$$\min_Q H(Q) \quad \text{subject to (5.11) and (5.12),} \tag{5.13}$$

which is similar to (5.6). We can show that the solution of (5.13) is given by

$$Q^\alpha = I^{\alpha\beta} I^{\gamma\delta} Z_{\beta\gamma} Z_\delta + \tfrac{1}{2} I^{\alpha\eta} I^{\mu\beta} I^{\kappa\gamma} R_{\eta\mu\kappa} Z_\beta Z_\gamma, \qquad \alpha = 1, ..., d,$$

which are exactly the second order terms in the stochastic expansion of the maximum likelihood estimator of λ.

References

[1] Amari, S. I. (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, Vol. 28, Springer-Verlag, New York/Berlin.

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

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

[4] BHATTACHARYA, R. N., AND GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451.

[5] BHATTACHARYA, R. N., AND RAO, R. R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.

[6] COCHRAN, W. G., AND BLISS, C. I. (1946). Discriminant functions with covariance. *Ann. Math. Statist.* **19** 151–176,

[7] FUJIKOSHI, Y., AND KANAZAWA, M. (1976). The ML classification statistic in covariate discriminant analysis and its asymptotic expansions. In *Essays in Probability and Statistics (Ogawa Volume)*, pp. 305–320. Shinko-Tsusho, Tokyo.

[8] HAN, C. P. (1969). Distribution of discriminant function when covariance matrices are proportional. *Ann. Math. Statist.* **40** 979–985.

[9] HAN, C. P. (1970). Distribution of discriminant function in circular models. *Ann. Inst. Statist. Math.* **22** 117–125.

[10] LAWLEY, D. N., AND MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. Butterworths, London.

[11] MEMON, A. Z., AND OKAMOTO, M. (1970). The classification statistic W^* in covariate discriminant analysis. *Ann. Math. Statist.* **41** 1491–1499.

[12] MEMON, A. Z., AND OKAMOTO, M. (1971). Asymptotic expansion of the distribution of the Z statistic in discriminant analysis. *J. Multivariate Anal.* **1** 294–307.

[13] OKAMOTO, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Statist.* **34** 1286–1301; Correction, **39** 1358–1359.

[14] SIOTANI, M. (1982). Large sample approximations and asymptotic expansions of classification statistics. In *Handbook of Statistics*, Vol. 2 (Krishnaiah and Kanal, Eds.), pp. 61–100. North-Holland, Amsterdam.

[15] SIOTANI, M., AND WANG, R. H. (1977). Asymptotic expansions for error rates and comparison of the W-procedure and the Z-procedure in discriminant analysis. In *Multivariate Analysis-IV* (P. R. Krishnaiah, Ed.), pp. 523–545. North-Holland, Amsterdam.

[16] TAKEUCHI, K., AND MORIMUNE, K. (1985). Third-order efficiency of the extended maximum likelihood estimators in a simultaneous equation system. *Econometrica* **53** 177–200.

[17] TANIGUCHI, M. (1991). *Higher Order Asymptotic Theory for Time Series Analysis*. Lecture Notes in Statistics, Vol. 68. Springer-Verlag, Heidelberg.

[18] WAKAKI, H. (1990). Comparison of linear and quadratic discriminant functions. *Biometrika* **77** 227–229.