Open Journal of Statistics, 2016, 6, 172-185
Published Online February 2016 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2016.61015

A Comparison of Two Linear Discriminant Analysis Methods That Use Block Monotone Missing Training Data

Phil D. Young[1], Dean M. Young[2], Songthip T. Ounpraseuth[3]

[1]Department of Information Systems, Baylor University, Waco, TX, USA
[2]Department of Statistical Science, Baylor University, Waco, TX, USA
[3]Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, AK, USA
Email: philip_young@baylor.edu, dean_young@baylor.edu, STOunpraseuth@uams.edu

Received 7 January 2016; accepted 23 February 2016; published 26 February 2016

Abstract

We revisit a comparison of two discriminant analysis procedures, namely the linear combination classifier of Chung and Han (2000) and the maximum likelihood estimation substitution classifier for the problem of classifying unlabeled multivariate normal observations with equal covariance matrices into one of two classes. Both classes have matching block monotone missing training data. Here, we demonstrate that for intra-class covariance structures with at least small correlation among the variables with missing data and the variables without block missing data, the maximum likelihood estimation substitution classifier outperforms the Chung and Han (2000) classifier regardless of the percent of missing observations. Specifically, we examine the differences in the estimated expected error rates for these classifiers using a Monte Carlo simulation, and we compare the two classifiers using two real data sets with monotone missing data via parametric bootstrap simulations. Our results contradict the conclusions of Chung and Han (2000) that their linear combination classifier is superior to the *MLE* classifier for block monotone missing multivariate normal data.

Keywords

Linear Discriminant Analysis, Monte Carlo Simulation, Maximum Likelihood Estimator, Expected Error Rate, Conditional Error Rate

1. Introduction

We consider the problem of classifying an unlabeled observation vector $x \sim N_p(\mu_i, \Sigma)$ into one of two distinct

How to cite this paper: Young, P.D., Young, D.M. and Ounpraseuth, S.T. (2016) A Comparison of Two Linear Discriminant Analysis Methods That Use Block Monotone Missing Training Data. *Open Journal of Statistics*, **6**, 172-185.
http://dx.doi.org/10.4236/ojs.2016.61015

multivariate normally distributed populations $\Pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, when monotone missing training data are present, where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$ are the i^{th} population mean vector and common covariance matrix, respectively. Here, we re-compare two linear classification procedures for block monotone missing (*BMM*) training data: one classifier is from [1], and the other classifier employs the maximum likelihood estimator (*MLE*).

Monotone missing data occur for an observation vector $\boldsymbol{x}_j$ when, if $\boldsymbol{x}_{ji}$ is missing, then $\boldsymbol{x}_{jk}$ is missing for all $k > i$. The authors [1] claim that their "linear combination classification procedure is better than the substitution methods (*MLE*) as the proportion of missing observations gets larger" when block monotone missing data are present in the training data. Specifically, [1] has performed a Monte Carlo simulation and has concluded that their classifier performs better in terms of the expected error rate (*EER*) than the *MLE* substitution (*MLES*) classifier formulated by [2] as the proportion of missing observations increases. However, we demonstrate that for intra-class covariance training data with at least small correlations among the variables, the *MLES* classifier can significantly outperform the classifier from [1], which we refer to as the *C-H* classifier, in terms of their respective *EER*s. This phenomenon occurs regardless of the proportion of the variables missing in each observation with missing data (*POMD*) in the training data set.

Throughout the remainder of the paper, we use the notation $\mathbb{R}_{m\times n}$ to represent the matrix space of all $m \times n$ matrices over the real field $\mathbb{R}$. Also, we let the symbol $\mathbb{R}_n^{>}$ represent the cone of all $n \times n$ positive definite matrices in $\mathbb{R}_{n\times n}$. Moreover, $\boldsymbol{A}' \in \mathbb{R}_{n\times m}$ represents the transpose of $\boldsymbol{A} \in \mathbb{R}_{m\times n}$.

The author [3] has considered the problem of missing values in discriminant analysis where the dimension and the training-sample sizes are very large. Additionally, [4] has examined the probability of correct classification for several methods of handling data values that are missing at random and use the *EER* as the criterion to weigh the relative quality of supervised classification methods. Moreover, [5] has examined missing observations in statistical discrimination for a variety of population covariance matrices. Also, [6] has applied recursive methods for handling incomplete data and has verified asymptotic properties for the recursive methods.

We have organized the remainder of the paper as follows. In Section 2, we describe the *C-H* classifier, and we describe the *MLES* linear discriminant procedure when the training data from both classes contain identical *BMM* data patterns. In Section 3, we describe and report the result of Monte Carlo simulations that examine the differences in the estimated *EER*s of the *C-H* and *MLES* classifiers for various parameter configurations, training-sample sizes, and missing data sizes and summarize our simulation results graphically. In Section 4, we compare the *C-H* and *MLES* linear classifiers using a parametric bootstrap estimator of the *EER* difference (*EERD*) on two actual data sets. We summarize our results and conclude with some brief comments in Section 5.

2. Two Competing Classifiers for *BMM* Training Data

2.1. The *C-H* Classifier for Monotone Missing Data

Suppose we have two $p \times N_i$ training observation matrices in the form

$$\begin{bmatrix} \boldsymbol{Y}_{i1} & \boldsymbol{Y}_{i2} \\ \boldsymbol{Z}_i & \cdot \end{bmatrix}, \tag{1}$$

where

$$\boldsymbol{U}_i = \left[\boldsymbol{Y}_{i1}' \ : \ \boldsymbol{Z}_i'\right]' \in \mathbb{R}_{k\times n_i} \tag{2}$$

denotes the n_i complete-observation submatrix, and $\boldsymbol{Y}_{i2} \in \mathbb{R}_{k\times(N_i - n_i)}$ is the partial observation submatrix whose first k measurements are non-missing, where $N_i > n_i$, for $i = 1, 2$. We denote a complete data observation vector by $\boldsymbol{u}_{ij} = \left[\boldsymbol{y}_{i1j}' : \boldsymbol{z}_{ij}'\right]'$, where $\boldsymbol{y}_{i1j} \in \mathbb{R}_{k\times 1}$ and $\boldsymbol{z}_{ij} \in \mathbb{R}_{(p-k)\times 1}$ such that

$$\boldsymbol{u}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}) \equiv N_p\left(\begin{bmatrix} \boldsymbol{\mu}_{Y_{i1}} \\ \boldsymbol{\mu}_{Z_i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right), \tag{3}$$

where $\boldsymbol{\mu}_{Y_{i1}} \in \mathbb{R}_{k\times 1}$, $\boldsymbol{\mu}_{Z_i} \in \mathbb{R}_{(p-k)\times 1}$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}_{k\times k}^{>}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}_{k\times(p-k)}$, and $\boldsymbol{\Sigma}_{22} \in \mathbb{R}_{(p-k)\times(p-k)}^{>}$ with $i = 1, 2;\ j = 1, 2, \cdots, n_i$. Also, random samples of sizes $N_i - n_i$ are taken from distributions of the form $N_k\left(\boldsymbol{\mu}_{Y_{i2}}, \boldsymbol{\Sigma}_{yy}\right)$, where $\boldsymbol{\mu}_{Y_{i2}} \in \mathbb{R}_{k\times 1}$ and $\boldsymbol{\Sigma}_{yy} \in \mathbb{R}_{k\times k}^{>}$.

The authors [1] have derived a linear combination of a discriminant function composed from complete data and a second discriminant function determined from *BMM* data. The *C-H* classifier uses Anderson's linear dis-

criminant function (*LDF*) for the subset of complete data $\boldsymbol{U}_i$, $i=1,2$, given in (2), which is

$$W_{\boldsymbol{u}} \equiv \left(\bar{\boldsymbol{u}}_1 - \bar{\boldsymbol{u}}_2\right)' \boldsymbol{S}_{\boldsymbol{u}}^{-1}\left[\boldsymbol{u} - \frac{1}{2}\left(\bar{\boldsymbol{u}}_1 + \bar{\boldsymbol{u}}_2\right)\right],$$

where $\bar{\boldsymbol{u}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{u}_{ij}$ and

$$\boldsymbol{S}_{\boldsymbol{u}} = \frac{1}{n_1 + n_2 - 2}\left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\boldsymbol{u}_{ij} - \bar{\boldsymbol{u}}_i\right)\left(\boldsymbol{u}_{ij} - \bar{\boldsymbol{u}}_i\right)'\right),$$

are the complete-data sample mean and complete-data sample covariance matrix, respectively. They also use Anderson's *LDF* for the data

$$\left[\boldsymbol{Y}_{i1} : \boldsymbol{Y}_{i2}\right], \tag{4}$$

$i=1,2$, with k features and $N_1 + N_2$ training observations, which is

$$W_{\boldsymbol{y}} \equiv \left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2\right)' \boldsymbol{S}_{\boldsymbol{y}}^{-1}\left[\boldsymbol{y} - \frac{1}{2}\left(\bar{\boldsymbol{y}}_1 + \bar{\boldsymbol{y}}_2\right)\right]$$

with

$$\bar{\boldsymbol{y}}_i = \frac{1}{N_i}\left[n_i \bar{\boldsymbol{y}}_{i1} + \left(N_i - n_i\right)\bar{\boldsymbol{y}}_{i2}\right], \tag{5}$$

where

$$\bar{\boldsymbol{y}}_{i1} = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{y}_{i1j} \tag{6}$$

denotes the sample mean for the first n_i observations and the first k features from $\boldsymbol{Y}_{i1}$ in (1),

$$\bar{\boldsymbol{y}}_{i2} = \frac{1}{N_i - n_i}\sum_{j=n_i+1}^{N_i} \boldsymbol{y}_{i2j}$$

denotes the sample mean for the first k features of the latter $N_i - n_i$ observations from $\boldsymbol{Y}_{i2}$ in (1), and

$$\boldsymbol{S}_{\boldsymbol{y}} = \frac{1}{N_1 + N_2 - 2}\left(\sum_{i=1}^{2}\sum_{t=1}^{2}\sum_{j=1}^{N_i}\left(\boldsymbol{y}_{itj} - \bar{\boldsymbol{y}}_i\right)\left(\boldsymbol{y}_{itj} - \bar{\boldsymbol{y}}_i\right)'\right)$$

is the pooled sample covariance matrix for the incomplete training data (4), where $t=1,2$, represent the subsets of (1) with non-missing data and *BMM* data, respectively, for $i=1,2$.

The authors [1] have proposed the linear combination statistic

$$W_c \equiv cW_{\boldsymbol{u}} + \left(1-c\right)W_{\boldsymbol{y}}, \tag{7}$$

where $c \in [0,1]$. One classifies an unlabeled observation vector $x \in \mathbb{R}_{p\times 1}$ into Π_1 if

$$W_c \geq 0 \tag{8}$$

and into Π_2, otherwise. The conditional error rate (*CER*) for classifying an unlabeled vector x from Π_i into Π_j using (8) is

$$\begin{aligned} CER_{ij}\left(W_c\right) &= P\left(\left(-1\right)^{2-j} W_c < 0 \,|\, \bar{\boldsymbol{u}}_1, \bar{\boldsymbol{u}}_2, \boldsymbol{S}_u, \bar{\boldsymbol{y}}_1, \bar{\boldsymbol{y}}_2, \boldsymbol{S}_y; \boldsymbol{u}, \boldsymbol{y} \in \Pi_i\right) \\ &= \Phi\left(\frac{\left(-1\right)^{2-i}\boldsymbol{h}'\boldsymbol{\mu}_i + \left(-1\right)^{2-i} f}{\sqrt{\boldsymbol{h}'\boldsymbol{\Sigma}\boldsymbol{h}}}\right), \end{aligned} \tag{9}$$

$i, j = 1,2$, $i \neq j$, where

$$f \equiv cb + \left(1-c\right)e \tag{10}$$

with

$$b \equiv -\frac{1}{2}\left(\bar{\boldsymbol{u}}_1 - \bar{\boldsymbol{u}}_2\right)' \boldsymbol{S}_u^{-1}\left(\bar{\boldsymbol{u}}_1 + \bar{\boldsymbol{u}}_2\right) \text{ and } e \equiv -\frac{1}{2}\left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2\right)' \boldsymbol{S}_y^{-1}\left(\bar{\boldsymbol{y}}_1 + \bar{\boldsymbol{y}}_2\right).$$

Also,

$$\boldsymbol{h} = \left[\boldsymbol{r}' \ : \ \boldsymbol{b}'\right]', \tag{11}$$

where $\boldsymbol{r} \in \mathbb{R}_{k\times 1}$ and $\boldsymbol{b} \in \mathbb{R}_{(p-k)\times 1}$, $\boldsymbol{r} \equiv c\boldsymbol{a}_1 + (1-c)\boldsymbol{d}$, $\boldsymbol{b} \equiv c\boldsymbol{a}_2$, $\boldsymbol{d} = \boldsymbol{S}_y^{-1}\left(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2\right)$, $\boldsymbol{a} = \boldsymbol{S}_u^{-1}\left(\bar{\boldsymbol{u}}_1 - \bar{\boldsymbol{u}}_2\right) = \left[\boldsymbol{a}_1' \ : \ \boldsymbol{a}_2'\right]$, with $\boldsymbol{a}_1 \in \mathbb{R}_{k\times 1}$, $\boldsymbol{a}_2 \in \mathbb{R}_{(p-k)\times 1}$, and $\bar{\boldsymbol{y}}_i$ defined in (5). Thus, using (9) and assuming equal *a priori* probabilities, the *CER* for (8) is

$$CER\left(W_c\right) = \frac{1}{2}\left[CER_{12}\left(W_c\right) + CER_{21}\left(W_c\right)\right]. \tag{12}$$

If

$$\tilde{\boldsymbol{\theta}} \equiv \left[\bar{\boldsymbol{y}}_1 : \bar{\boldsymbol{y}}_2 : \boldsymbol{S}_y : \boldsymbol{S}_u : \bar{\boldsymbol{u}}_1 : \bar{\boldsymbol{u}}_2\right],$$

then, for (8), the *EER* of misclassifying an unlabeled observation vector $\boldsymbol{x}$ from Π_i into Π_j is

$$EER\left(W_c\right)_{ij} = E_{\tilde{\theta}}\left[\Phi\left(\frac{(-1)^{2-i}\,\boldsymbol{h}'\boldsymbol{\mu}_i + (-1)^{2-i}\,f}{\boldsymbol{h}'\boldsymbol{\Sigma}\boldsymbol{h}}\right)\right],$$

$i, j = 1, 2$, $i \neq j$. Thus, once again assuming equal *a priori* probabilities, the *EER* for (8) is

$$EER\left(W_c\right) = \frac{1}{2}\left[EER_{12}\left(W_c\right) + EER_{21}\left(W_c\right)\right].$$

In choosing c in (7), [1] have utilized the fact that the *CER* and *EER* will depend on the Mahalanobis distance for the complete and partial training observations and the corresponding training-sample sizes, N_i and n_i, $i = 1, 2$. Usually, when one has small *CER*s, at least one of the sample Mahalanobis distances

$$D_w^2 \equiv \left(\bar{\boldsymbol{w}}_1 - \bar{\boldsymbol{w}}_2\right)' \boldsymbol{S}_w^{-1}\left(\bar{\boldsymbol{w}}_1 - \bar{\boldsymbol{w}}_2\right), \quad \boldsymbol{w} = \boldsymbol{u}, \boldsymbol{y},$$

will be large. While n_i and $D_{\boldsymbol{u}}^2$ determine the performance of $W_{\boldsymbol{u}}$, the quantities N_i and $D_{\boldsymbol{y}}^2$ dictate the performance of $W_{\boldsymbol{y}}$. Hence, [1] have chosen c in relation to the training-sample sizes and the Mahalanobis distances for the complete and incomplete training-data sets. Note that the implication for circumstances where $D_{\boldsymbol{u}}^2 > D_{\boldsymbol{y}}^2$ is that the information in the data-matrix component $\boldsymbol{Z}_i$, $i = 1, 2$, in (1) contributes largely to the discriminatory information. Hence, [1] uses

$$c^* = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} D_{\boldsymbol{u}}^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} D_{\boldsymbol{u}}^2 + \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} D_{\boldsymbol{y}}^2}$$

to determine the linear combination classification statistic (7).

2.2. A Maximum Likelihood Substitution Classifier for Monotone Missing Training Data

The authors [7] have derived an *MLE* method for estimating parameters in a multivariate normal distribution with *BMM* data. The estimator of $\boldsymbol{\Sigma}$ in the [7] *MLES* classifier is a pooled estimator of the two individual *MLE*s of $\boldsymbol{\Sigma}$.

Below, we state the *MLE*s for two multivariate normal distributions having unequal means and a common covariance matrix with identical *BMM*-data patterns in both training samples.

Theorem. Let Π_i be modeled by the multivariate normal densities $N_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$ for $i = 1, 2$, with

$$\boldsymbol{\mu}_i \equiv \begin{bmatrix} \boldsymbol{\mu}_{i1} \\ \boldsymbol{\mu}_{i2} \end{bmatrix} \tag{13}$$

and

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}. \tag{14}$$

Also, let

$$\boldsymbol{A}_{11,N_i,i} \equiv \sum_{j=1}^{N_i} \left(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i\right)\left(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i\right)',$$

$$\boldsymbol{A}_{11,n_i,i} \equiv \sum_{j=1}^{n_i} \left(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i\right)\left(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i\right)',$$

$$\boldsymbol{A}_{12,n_i,i} \equiv \sum_{j=1}^{n_i} \left(\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_i\right)\left(\boldsymbol{z}_{ij} - \overline{\boldsymbol{z}}_i\right)',$$

and

$$\boldsymbol{A}_{22,n_i,i} \equiv \sum_{j=1}^{n_i} \left(\boldsymbol{z}_{ij} - \overline{\boldsymbol{z}}_i\right)\left(\boldsymbol{z}_{ij} - \overline{\boldsymbol{z}}_i\right)',$$

where $\boldsymbol{y}_{ij} \in \left[\boldsymbol{Y}_{i1} : \boldsymbol{Y}_{i2}\right]$, and $\boldsymbol{z}_{ij} \in \boldsymbol{Z}_i$ with $\boldsymbol{Y}_{i1}$, $\boldsymbol{Y}_{i2}$, and $\boldsymbol{Z}_i$ given in (1). Then, on the basis of two-step monotone training samples from populations $\Pi_i : N_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right), i = 1, 2$, the *MLE*s of (13) and (14) are

$$\hat{\boldsymbol{\mu}}_i \equiv \begin{bmatrix} \hat{\boldsymbol{\mu}}_{i1} \\ \hat{\boldsymbol{\mu}}_{i2} \end{bmatrix} \text{ and } \hat{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{11} & \hat{\boldsymbol{\Sigma}}_{12} \\ \hat{\boldsymbol{\Sigma}}_{21} & \hat{\boldsymbol{\Sigma}}_{22} \end{bmatrix}, \tag{15}$$

respectively, where

$$\hat{\boldsymbol{\Sigma}}_{11} \equiv \frac{\sum_{i=1}^{2} \boldsymbol{A}_{11,N_i,i}}{\sum_{i=1}^{2} N_i},$$

$$\hat{\boldsymbol{\Sigma}}_{12} \equiv \frac{1}{\left(\sum_{i=1}^{2} N_i\right)} \left[\sum_{i=1}^{2} \boldsymbol{A}_{11,N_i,i}\right] \left[\sum_{i=1}^{2} \boldsymbol{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \boldsymbol{A}_{12,n_i,i}\right], \tag{16}$$

and

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{22} \equiv & \frac{1}{\left(\sum_{i=1}^{2} n_i\right)} \sum_{i=1}^{2} \boldsymbol{A}_{22\cdot 1,n_i,i} + \frac{1}{\left(\sum_{i=1}^{2} N_i\right)} \left[\sum_{i=1}^{2} \boldsymbol{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \boldsymbol{A}_{11,n_i,i}\right]^{-1} \\ & \times \left[\sum_{i=1}^{2} \boldsymbol{A}_{11,N_i,i}\right] \left[\sum_{i=1}^{2} \boldsymbol{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \boldsymbol{A}_{12,n_i,i}\right], \end{aligned} \tag{17}$$

with $\hat{\boldsymbol{\mu}}_{i1} = \overline{\boldsymbol{y}}_i$, where $\overline{\boldsymbol{y}}_i$ is defined in (5),

$$\hat{\boldsymbol{\mu}}_{i2} \equiv \overline{\boldsymbol{z}}_i - \left[\hat{\boldsymbol{\Sigma}}_{12} \hat{\boldsymbol{\Sigma}}_{22}^{-1}\right] \left(\overline{\boldsymbol{y}}_{i1} - \overline{\boldsymbol{y}}_{i2}\right),$$

$$\overline{\boldsymbol{z}}_i \equiv \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{z}_{ij},$$

and

$$\sum_{i=1}^{2} \boldsymbol{A}_{22\cdot 1,n_i,i} \equiv \sum_{i=1}^{2} \boldsymbol{A}_{22,n_i,i} - \left[\sum_{i=1}^{2} \boldsymbol{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \boldsymbol{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \boldsymbol{A}_{12,n_i,i}\right],$$

where $\bar{y}_{i1}$, $\bar{y}_{i2}$, $\hat{\boldsymbol{\Sigma}}_{12}$, and $\hat{\boldsymbol{\Sigma}}_{22}$, are defined in (5), (6), (16), and (17), respectively, for $i = 1, 2$.

Proof: A proof is alluded to in [8].

The *MLES* classification statistic is

$$W_{MLE} \equiv \left(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1\right)' \hat{\boldsymbol{\Sigma}}^{-1} \left[\boldsymbol{x} - \frac{1}{2}\left(\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\right)\right], \tag{18}$$

where $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$ are the *MLE*s defined in (15), and $\boldsymbol{x} \in \mathbb{R}_{p\times 1}$ is an unlabeled observation vector belonging to either Π_1 or Π_2. We classify the unlabeled observation vector $\boldsymbol{x} \in \mathbb{R}_{p\times 1}$ into Π_1 if

$$W_{MLE} \le 0 \tag{19}$$

and into Π_2, otherwise. Given that $\boldsymbol{x} \in \Pi_1$, conditioning on $\hat{\boldsymbol{\mu}}_{ij}$, $i, j = 1, 2$, and $\hat{\boldsymbol{\Sigma}}$, and using the fact that

$$\hat{\delta}'\hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}_1\right) \sim N\left(0, \hat{\delta}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\delta}\right),$$

where $\hat{\delta} \equiv \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$, along with (15), (18), and (19), we have that

$$CER_{12}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}\right) \equiv P\left[W_{MLE} > 0 \mid \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}; \boldsymbol{x} \in \Pi_1\right] = 1 - \Phi\left(w_1\right),$$

where

$$w_i \equiv \left[\hat{\delta}'\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\delta}\right]^{-1/2}\left\{\hat{\delta}'\hat{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\right) - \mu_i\right)\right\}, \tag{20}$$

$i = 1, 2$. Similarly, given $\boldsymbol{x} \in \Pi_2$,

$$CER_{21}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}\right) \equiv P\left[W_{MLE} \le 0 \mid \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}; \boldsymbol{x} \in \Pi_2\right] = \Phi\left(w_2\right),$$

where w_2 is given in (20). Thus, assuming equal *a priori* probabilities of belonging to Π_i, $i = 1, 2$, for an unlabeled observation, we have

$$CER\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}\right) \equiv \frac{1}{2}\left[1 - \Phi\left(w_1\right) + \Phi\left(w_2\right)\right]. \tag{21}$$

Hence, the overall expected error rate is

$$EER\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \hat{\boldsymbol{\Sigma}}\right) \equiv \frac{1}{2}\left[1 - E_{\tilde{\theta}}\left(\Phi\left(w_1\right)\right) + E_{\tilde{\theta}}\left(\Phi\left(w_2\right)\right)\right].$$

3. Monte Carlo Simulations

The authors [1] claim that "it can be shown that the linear combination classification statistic is invariant under nonsingular linear transformations when the data contain missing observations" and assume this invariance is also true for the *MLES* classifier. While their assertion might be true for the *C-H* classifier, it is not necessarily true for the *MLE* classifier. Because [1] do not consider covariance structures with moderate to high correlation, their results are biased toward the *C-H* classifier. Here, we show that the *MLES* classifier can considerably outperform the *C-H* classifier, depending on the degree of correlation among the variables with missing data and the variables without missing data.

Next, we present a description and results of a Monte Carlo simulation we have performed to evaluate the *EERD* between the *MLE* and *C-H* classifiers for two multivariate normal configurations, $\Pi_i : N_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$, $i = 1, 2$, using various training-sample sizes, dimensions, features with block missing data, differences in means, values of correlation among variables, and missing-data proportions. For the simulations, we define p to be the total number of feature dimensions and r to be the number of missing features so that $r < p$. Also, N_i denotes the total training-sample size from population Π_i, $i = 1, 2$, and

$$\boldsymbol{\Sigma} \equiv \rho\boldsymbol{J} + \left(1 - \rho\right)\boldsymbol{I}$$

is the intraclass covariance matrix where ρ denotes the common population correlation among the features in the intraclass covariance matrix and $J \in \mathbb{R}_{p\times p}$ denotes a matrix of ones.

The simulation was performed in SAS 9.2 (SAS Institute In., Cary, NC, USA) using the RANDNORMAL

command in PROC IML to generate 10,000 training-sample sets of size N_i, $i=1,2$, for each parameter configuration. Next, the *MLE* and *C-H* classifiers were computed, and their *CER*s were calculated for each training-sample set. Then, the differences between the *CER*s for the classifiers were averaged over the 10,000 *CER* differences for the two classifiers for each parameter configuration involving N_i, p, r, $\boldsymbol{\Sigma}$, $\boldsymbol{\mu}_i$, and *POMD* for the r features with monotone missing data, where $i=1,2$. Thus, the $\widehat{EERD}$ for the *C-H* and *MLES* classifiers is

$$\widehat{EERD}=\frac{1}{K}\sum_{j=1}^{K}\left[CER_j\left(W_c\right)-CER_j\left(\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2,\hat{\boldsymbol{\Sigma}}\right)\right],$$

where $CER\left(W_c\right)$ is defined in (12), $CER\left(\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2,\hat{\boldsymbol{\Sigma}}\right)$ is given in (21), K is the total number of simulated training-data sets, and j denotes the j^{th} simulated training-data set, where $j\in\{1,2,\cdots,K\}$. We display the results of our two Monte Carlo simulations by graphing $\widehat{EERD}$ against $\rho=0.1,0.3,0.5,0.7,0.9$ for various configurations of p, r, N_i, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and *POMD*.

The relationship between p and r was fixed at $r=0.2p$ and $r=0.8p$. We chose these specific values of p and r to evaluate $\widehat{EERD}$ when the proportion of variables with missing data were both small and large relative to p. The choice of r and N_i depended on the value of p, and we provide the values of p, r, and N_i used in the Monte Carlo simulation in **Table 1**.

Lastly, we chose $\boldsymbol{\mu}_1\in\mathbb{R}_{p\times 1}$ such that

$$\boldsymbol{\mu}_1=\left[0,0,\cdots,0\right]' \tag{22}$$

and $\boldsymbol{\mu}_2\in\mathbb{R}_{p\times 1}$ such that

$$\boldsymbol{\mu}_2=\left[d_j,0,0,\cdots,0,d_j,0,\cdots,0\right]', \tag{23}$$

with $d_1=0.5$ and $d_2=3$ to assess $\widehat{EERD}$ for both small and large between-class separation. These values for $\boldsymbol{\mu}_i$, $i=1,2$, given in (22) and (23), were chosen because they are similar to the population means used in the simulation used in [1]. Furthermore, we contrasted (8) and (19) using *POMD* = 0.5, 0.8 for the r covariates with *BMM* data, and as in [1], we chose $N_i>p$ to avoid singularity of the estimated covariance matrices. The comparison criterion $\widehat{EERD}$ is plotted against ρ for various combinations of p, r, d_j, N_i, and *POMD* in **Figure 1** and **Figure 2**, $i,j=1,2$. Although we simulated values for $\widehat{EERD}$ for $p=10,20,40$, we omitted the graphs for $p=20$ because the graphs are similar to the plots for $p=10$ and $p=40$. The graphs for $p=20$ can be obtained from the authors.

Figure 1, **Figure 2** illustrate that the $\widehat{EERD}$ is consistently positive for the values of p, r, N_i, ρ, d_j, and *POMD* examined here. Moreover, the figures indicate that the primary parameters that influence the dominance of the *MLES* classifier are ρ and d_j, $j=1,2$. For all feature dimensions considered here, the *C-H* and *MLES* classifiers were competitive for $\rho=0.1$. More importantly, for $\rho>0.1$, $\widehat{EERD}$ increased as ρ increased for all p, r, N_i, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and *POMD* considered here. The most noteworthy increase in the $\widehat{EERD}$ was for $0.7\le\rho\le 0.9$ when $d_1=0.5$, where $\widehat{EERD}$ increased by approximately 0.10. This increase occurred for all specified values of p, r, N_i, and *POMD*, and, thus, supported the superiority of the *MLES* classifier in terms of *EERD* for these configurations. Additionally, we noted that $\widehat{EERD}\approx 0.20$ when $\rho=0.9$, $d_1=0.5$, and other parameters are allowed to vary.

The *MLES* classifier especially outperformed the *C-H* classifier when $d_1=0.5$ for $\rho>0$, as compared to when $d_2=3$. The smaller values of $\widehat{EERD}$ for $d_2=3$ can be attributed to the fact that for a relatively large

Table 1. Dimensions and sample sizes for the Monte Carlo simulation.

p	r	N_i
10	2,8	20, 50, 100
20	4,16	25, 50, 100
40	8,32	50, 100, 200

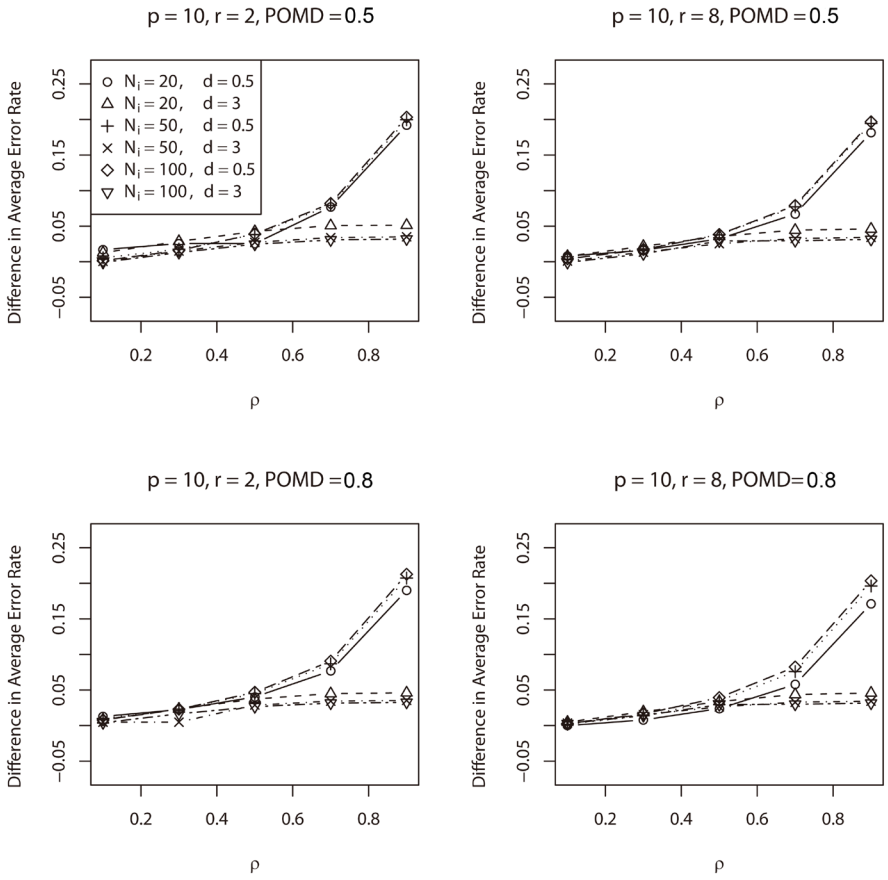

Figure 1. Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, *POMD*, and $p = 10$.

Mahalanobis distance when $d_2 = 3$ and $\rho = 0.1$, the *EER*s for both classifiers are small, thus yielding a smaller $\widehat{EERD}$.

As we used a large number of simulation iterations, we obtained $\max_{\boldsymbol{\xi}} \left\{ \widehat{s.e.}\left(\widehat{EERD} \right) \right\} < 0.003$, where $\boldsymbol{\xi}$ is the grid of parameter vectors considered in the simulation. Thus, the relatively small estimated standard errors also support our claim that $EERD > 0$ for $\rho > 0.1$ for the parameter configurations considered here. As **Figure 1**, **Figure 2** indicate, the contrasting values of p, r, N_i, and *POMD* contribute marginally, if at all, to $\widehat{EERD}$. Regardless of the combination of parameter values considered here, the *MLES* classifier dominates the *C-H* classifier in terms of $\widehat{EERD}$.

In summary, the simulation results indicated that the *MLES* classifier became increasingly superior to the *C-H* classifier as the correlation magnitude among the features with no missing data and the features with *BMM* data increased.

We remark that the standard errors for the $\widehat{EERD}$ in the [1] simulations are not sufficiently small enough to conclude a difference in the *ERR*s of the two competing classifiers. Hence, their claim that the *C-H* classifier outperforms the *MLES* classifier as the percent of missing observations increases is questionable.

We also performed a second Monte Carlo simulation whose results are not presented here. In this simulation, all fixed parameter values were equivalent to those of the first simulation except for $\boldsymbol{\mu}_2$ in (23), where we chose 0.80 of the elements of $\boldsymbol{\mu}_2$ to be non-zero. Consequently, we obtained slightly different results from those of our first simulation. However, the *MLES* classifier still outperformed the *C-H* classifier for all parameter configurations when $\rho \geq 0.1$. These results suggest that for classification problems with equal intra-class covariance matrices the *MLES* classifier is superior to the *C-H* classifier when at least small correlation exists among the features with missing data and the features without missing data.

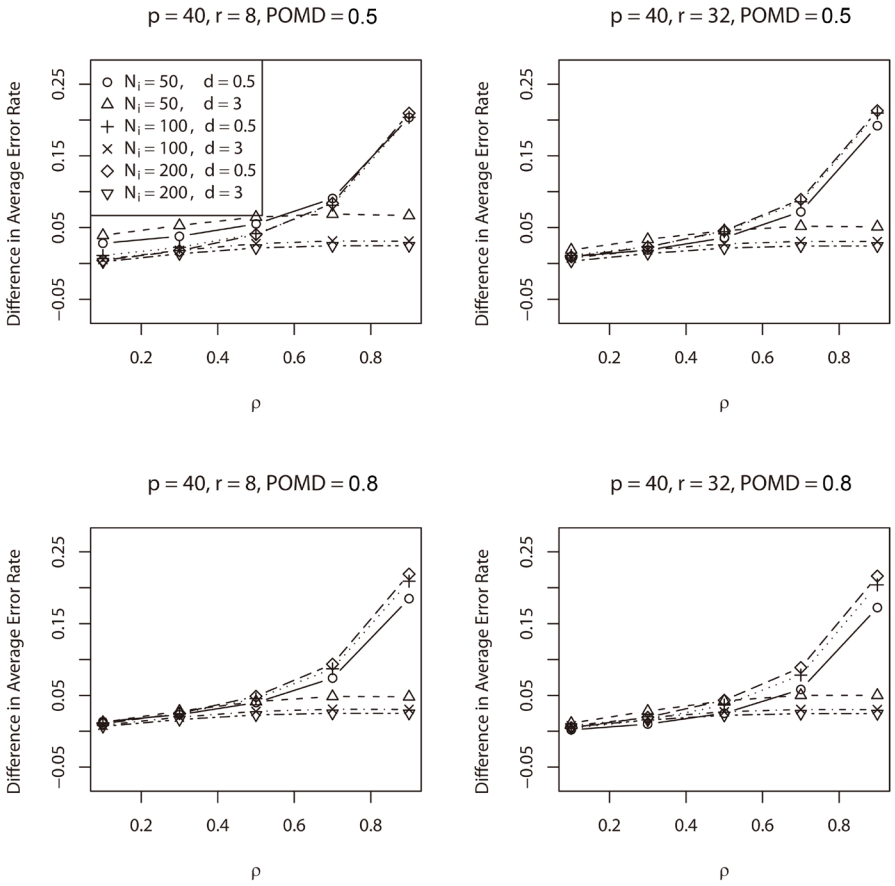

Figure 2. Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, *POMD*, and $p = 40$.

4. Two Real-Data Examples

4.1. Bootstrap Expected Error Rate Estimators for the *C-H* and *MLE* Classifiers

In this section, we compare the parametric bootstrap estimated *ERR*s of the *C-H* and *MLES* classifiers for two real-data sets each having two approximate multivariate normal populations with different population means and equal covariance matrices. First, we define the bootstrap *ERR* estimator for the *C-H* classifier, $\widehat{EER}_{\text{Boot}(C\text{-}H)}$. Let $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$ be the *MLE*s of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$, respectively, defined in Theorem 1. Also, let $\hat{\boldsymbol{\mu}}_1^*$, $\hat{\boldsymbol{\mu}}_2^*$, and $\hat{\boldsymbol{\Sigma}}^*$ be the bootstrap estimates of $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$, respectively, calculated using the parametric bootstrap training-sample data

$$\begin{bmatrix} \mathbf{Y}_{i1}^* & \mathbf{Y}_{i2}^* \\ \mathbf{Z}_i^* & \cdot \end{bmatrix} \tag{24}$$

that is generated from $N_p\left(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}\right)$, $i = 1, 2$. Then, conditioning on $\hat{\boldsymbol{\mu}}_i^*$, $i = 1, 2$, and $\hat{\boldsymbol{\Sigma}}^*$, the bootstrap *CER*s for the *C-H* classifier are

$$CER_{ij}^*\left(W_c^*\right) \equiv \Phi\left(\frac{(-1)^{2-i}\,\boldsymbol{h}^{*\prime}\hat{\boldsymbol{\mu}}_i + (-1)^{2-i}\,f^*}{\sqrt{\boldsymbol{h}^{*\prime}\hat{\boldsymbol{\Sigma}}\boldsymbol{h}^*}}\right)$$

for $i, j = 1, 2$, $i \neq j$, where W_c^*, $\boldsymbol{h}^*$, and f^* are similar in definition to W_c, $\boldsymbol{h}$, and f in (7), (11), and (10), respectively, except that we use the bootstrap multivariate normal data in (24). Thus, assuming equal prior probabilities, the bootstrap *CER* for the *C-H* classifier is

$$CER^*\left(W_c^*\right) \equiv \frac{1}{2}\left[CER_{12}^*\left(W_c^*\right) + CER_{21}^*\left(W_c^*\right)\right]. \tag{25}$$

Also, conditioning on $\hat{\boldsymbol{\mu}}_i^*$, $i = 1, 2$, and $\hat{\boldsymbol{\Sigma}}^*$, the bootstrap *CER*s for the *MLES* classifier are

$$CER_{ij}^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \hat{\boldsymbol{\Sigma}}^*\right) \equiv P\left[\left(-1\right)^{2-j} W_{MLE}^* > 0 \,|\, \hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \hat{\boldsymbol{\Sigma}}^*; \boldsymbol{x} \in \Pi_i\right],$$

where $\boldsymbol{x}$ is a complete unlabeled observation from $\Pi_1 \bigcup \Pi_2$, W_{MLE}^* is similar in definition to W_{MLE} in (18), and $i, j = 1, 2$, $i \neq j$. Given $\boldsymbol{x} \in \Pi_1$ and $\hat{\delta}^* \equiv \hat{\boldsymbol{\mu}}_1^* - \hat{\boldsymbol{\mu}}_2^*$, we have

$$CER_{12}^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \hat{\boldsymbol{\Sigma}}^*\right) = 1 - \Phi\left(w_1^*\right),$$

and given $\boldsymbol{x} \in \Pi_2$,

$$CER_{21}^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \hat{\boldsymbol{\Sigma}}^*\right) = \Phi\left(w_2^*\right),$$

where

$$w_i^* \equiv \left[\hat{\delta}^{*\prime}\hat{\boldsymbol{\Sigma}}^{*-1}\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Sigma}}^{*-1}\hat{\delta}^*\right]^{-1/2}\left[\hat{\delta}^{*\prime}\hat{\boldsymbol{\Sigma}}^{*-1}\left(\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_2^* + \hat{\boldsymbol{\mu}}_1^*\right) - \hat{\boldsymbol{\mu}}_i\right)\right], i = 1, 2.$$

Thus, assuming equal *a priori* probabilities of belonging to Π_i, $i = 1, 2$, for an unlabeled observation, we have

$$CER^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \hat{\boldsymbol{\Sigma}}^*\right) = \frac{1}{2}\left[1 - \Phi\left(w_1^*\right) + \Phi\left(w_2^*\right)\right]. \tag{26}$$

Hence, the estimated parametric bootstrap *EERD* for the *C-H* and *MLES* classifiers is

$$\widehat{EERD}_{\text{Boot}} \equiv \frac{1}{K}\sum_{j=1}^{K}\left(\widehat{CER}_{j\text{Boot}(C\text{-}H)} - \widehat{CER}_{j\text{Boot}(MLE)}\right), \tag{27}$$

where j denotes the j^{th} simulated training-data set for $j \in \{1, 2, \cdots, K\}$. We use (27) to compare the *C-H* and *MLES* classifiers for two real-data sets given in the following subsections.

4.2. A Comparison of the *C-H* and *MLE* Classifiers for *UTA* Admissions Data

The first data set was supplied by the Admissions Office at the University of Texas at Arlington and implemented as an example in [1]. The two populations for the *UTA* data are the Success Group for the students who receive their master's degrees (Π_1) and the Failure Group for students who do not complete their master's degrees (Π_2). Each training sample is composed of ten foreign students and ten United States students. Each foreign student had 5 variables associated with him or her. The variables are X_1 = undergraduate GPA, X_2 = GRE verbal, X_3 = GRE quantitative, X_4 = GRE analytic, and X_5 = TOEFL score. For each observation in both data sets, variables X_1, X_2, X_3, and X_4 are complete; however, X_5 contains monotone missing data. The *UTA* data set as seen in [1] can be seen in **Table 2**.

Also, the common estimated correlation matrix for the *UTA* data is

$$\hat{\boldsymbol{C}}_{UTA} = \begin{bmatrix} 1.000 & 0.145 & -0.066 & 0.199 & 0.373 \\ 0.145 & 1.000 & -0.404 & 0.494 & 0.767 \\ -0.066 & -0.404 & 1.000 & 0.129 & -0.493 \\ 0.199 & 0.494 & 0.129 & 1.000 & 0.392 \\ 0.373 & 0.767 & -0.493 & 0.392 & 1.000 \end{bmatrix}. \tag{28}$$

We remark that only one sample correlation coefficient in the last column of (28) has a magnitude exceeding 0.50, which reflects relatively low correlation among the four features without *BMM* data with the one feature having *BMM* data.

To estimate *EERD* for the *C-H* classifier (8) and the *MLES* classifier (19) for the *UTA* Admissions data, we determine $\widehat{EERD}_{\text{Boot}}$, given in (27), using 10,000 bootstrap simulation iterations with $p = 5$, $r = 1$, $N_i = 20$,

Table 2. UTA Admissions office.

Π_1 : Success					Π_2 : Failure				
x_1	x_2	x_3	x_4	x_5	x_1	x_2	x_3	x_4	x_5
2.97	420	800	600	497	3.75	250	730	460	513
3.80	330	710	380	563	3.11	320	760	610	560
2.50	270	700	340	510	3.00	360	720	525	540
2.50	400	710	600	563	2.60	370	780	500	500
3.30	280	800	450	543	3.50	300	630	380	507
2.60	310	660	425	507	3.50	390	580	370	587
2.70	360	620	590	537	3.10	380	770	500	520
3.10	220	530	340	543	2.30	370	640	200	520
2.60	350	770	560	580	2.85	340	800	540	517
3.20	360	750	440	577	3.50	460	750	560	597
3.65	440	700	630		3.15	630	540	600	
3.56	640	520	610		2.93	350	690	620	
3.00	480	550	560		3.20	480	610	480	
3.18	550	630	630		2.76	630	410	530	
3.84	450	660	630		3.00	550	450	500	
3.18	410	410	340		3.28	510	690	730	
3.43	460	610	560		3.11	640	720	520	
3.52	580	580	610		3.42	440	580	620	
3.09	450	540	570		3.00	350	430	480	
3.70	420	630	660		2.67	480	700	670	

and $n_i = 10$ for $i = 1, 2$. Additionally, the parametric bootstrap multivariate normal distribution parameters, which are the *MLE*s for the multivariate normal population parameters given in Theorem 1, are

$$\hat{\boldsymbol{\mu}}_1 = \left[3.171, 409, 644, 526.25, 577.01\right]'$$

and

$$\hat{\boldsymbol{\mu}}_2 = \left[3.087, 430, 649, 519.75, 562.66\right]'$$

for the means of Π_1 and Π_2, respectively, with common covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.150 & 6.020 & -2.760 & 8.540 & 6.510 \\ 6.020 & 11504.500 & -4683 & 5859.375 & 3711.097 \\ -2.760 & -4683 & 11701.500 & 1518.625 & -2406.740 \\ 8.540 & 5859.375 & 1518.625 & 12229.187 & 1953.163 \\ 6.510 & 3711.097 & -2406.740 & 1953.163 & 2034.414 \end{bmatrix}.$$

Subsequently, we obtained $\widehat{EERD}_{\text{Boot}} = -0.027$ with $\widehat{s.e.}\left(\widehat{EERD}\right)_{\text{Boot}} = 0.001$, which indicated that the *C-H* classifier yielded slightly better discriminatory performance compared to the *MLES* classifier for the *UTA* data. The fact that the *C-H* procedure slightly outperformed the *MLES* classifier for the *UTA* data set in terms of *EERD* is not surprising. In the *UTA* data set, relatively little correlation exists among many of the features, and

the *C-H* classifier does not require or use information in the correlation between the features with no missing data and the features with missing data. However, the *MLES* classifier does require at least a moderate degree of correlation between some features with no missing data and the feature with missing data to yield a more effective supervised classifier than the *C-H* classifier.

4.3. A Comparison of the *C-H* and *MLE* Classifiers on the Partial Iris Data

The second real-data set on which we compare the *C-H* and *MLES* classifiers is a subset of the well-known Iris data, which is one of the most popular data sets applied in pattern recognition literature and was first analyzed by R. A. Fisher (1936). The data used here is given in **Table 3**.

The University of Irvine Machine Learning Repository website provides the original data set, which contains 150 observations (50 in each class) with four variables: X_1 = sepal length (cm), X_2 = sepal width (cm), X_3 = petal length (cm), and X_4 = petal width (cm). This data set has three classes: Iris-setosa (Π_1), Iris-versicolor (Π_2), and Iris-virginica (Π_3). We have used a subset of the original Iris data set by taking only the first 20 observations from Π_1 and Π_2 and omitting the Iris-virginica group (Π_3). We emphasize that the variables in the partial Iris data are much more highly correlated than the variables in the *UTA* data. The estimated correlation matrix is

$$\hat{\boldsymbol{C}}_{Iris} = \begin{bmatrix} 1 & 0.716 & 0.708 & 0.549 \\ 0.716 & 1 & 0.473 & 0.651 \\ 0.708 & 0.473 & 1 & 0.677 \\ 0.549 & 0.651 & 0.677 & 1 \end{bmatrix}. \tag{29}$$

Table 3. Partial iris data.

Π_1 : Setosa				Π_2 : Versicolor			
x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.6
5.4	3.7	1.5		5.0	2.0	3.5	
4.8	3.4	1.6		5.9	3.0	4.2	
4.8	3.0	1.4		6.0	2.2	4.0	
4.3	3.0	1.1		6.1	2.9	4.7	
5.8	4.0	1.2		5.6	2.9	3.6	
5.7	4.4	1.5		6.7	3.1	4.4	
5.4	3.9	1.3		5.6	3.0	4.5	
5.1	3.5	1.4		5.8	2.7	4.1	
5.7	3.8	1.7		6.2	2.2	4.5	
5.1	3.8	1.5		5.6	2.5	3.9	

In (29), all estimated correlation coefficients in the last column had a magnitude greater than 0.50, which reflects a moderate degree of correlation among the features X_1, X_2, and X_3, and the feature X_4, which has *BMM* data.

For the Iris subset data, which can be found in **Table 3**, we used 10,000 bootstrap iterations, $N_i = 20$, $n_i = 10$, $POMD = 0.50$, $p = 4$, and $r = 1$, where $i = 1, 2$, for calculating $\widehat{EERD}_{\text{Boot}}$. Hence, the overall proportion of missing observations for the Iris subset data is greater than that of the *UTA* data set. The bootstrap parameters corresponding to Π_1 and Π_2 are

$$\hat{\boldsymbol{\mu}}_1 = \left[5.035, 3.48, 1.435, 0.235\right]'$$

and

$$\hat{\boldsymbol{\mu}}_2 = \left[5.975, 2.76, 4.255, 1.325\right]',$$

respectively, with common covariance matrix

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.273 & 0.147 & 0.124 & 0.045 \\ 0.147 & 0.154 & 0.062 & 0.040 \\ 0.124 & 0.062 & 0.111 & 0.035 \\ 0.045 & 0.040 & 0.035 & 0.024 \end{bmatrix}.$$

For the parametric bootstrap estimate for $\widehat{EERD}$ corresponding to the *C-H* and *MLES* classifiers applied to the subset of the Iris data set, we obtained $\widehat{EERD}_{\text{Boot}} = 0.11$ with $\widehat{s.e.}\left(\widehat{EERD}_{\text{Boot}}\right) = 0.001$, which indicated that $EER_{(MLE)} \ll EER_{(C\text{-}H)}$. Consequently, because of the relatively large correlations among the variables with no missing data, namely, X_1, X_2, X_3, and the variable with missing data, X_4, the *MLES* classifier convincingly outperforms the *C-H* classifier in terms of *EERD*. This evidence essentially contradicts the conclusion in [1] that the *C-H* classifier is superior to the *MLES* classifier when the proportion of observations with missing data is substantial, regardless of the covariance structure.

5. Conclusions

In this paper, we have considered the problem of supervised classification using training data with identical *BMM* data patterns for two multivariate normal classes with unequal means and equal covariance matrices. In doing so, we have used a Monte Carlo simulation to demonstrate that for the various parameter configurations considered here, ρ, not *POMD*, has the greatest impact on *EERD*. We have also concluded that the *MLES* classifier outperforms the *C-H* classifier for all considered parameter configurations involving intra-class covariance structures when $\rho \geq 0.1$ and becomes an increasingly superior statistical classification procedure as ρ approaches 1. This conclusion essentially contradicts the simulation results of [1].

We also have compared the *MLE* and *C-H* classifiers on two real training-data sets using $\widehat{EERD}_{\text{Boot}}$ in (27). From the real data set in [1], we have demonstrated that the *C-H* classifier can perform slightly better than the *MLES* classifier. Moreover, we have used a subset of the prominent Iris data set from [9] to illustrate that when the magnitude of the correlation among features without missing data and features with missing data is moderate to large, the *MLES* classifier is superior to the *C-H* classifier.

References

[1] Chung, H.-C. and Han, C.-P. (2000) Discriminant Analysis When a Block of Observations Is Missing. *Annals of the Institute of Statistical Mathematics*, **52**, 544-556.

[2] Bohannon, T.R. and Smith, W.B. (1975) Classification Based on Incomplete Data Records. *ASA Proceeding of Social Statistics Section*, **67**, 214-218.

[3] Jackson, E.C. (1968) Missing Values in Linear Multiple Discriminant Analysis. *Biometrics*, **24**, 835-844. http://dx.doi.org/10.2307/2528874

[4] Chang, L.S. and Dunn, O.J. (1972) The Treatment of Missing Values in Discriminant Analysis—1. The Sampling Experiment. *Journal of the American Statistical Association*, **67**, 473-477.

[5] Chang, L.S., Gilman, A. and Dunn, O.J. (1976) Alternative Approaches to Missing Values in Discriminant Analysis.

Journal of the American Statistical Association, **71**, 842-844. http://dx.doi.org/10.1080/01621459.1976.10480956

[6] Titterington, D.M. and Jian, J.-M. (1983) Recursive Estimation Procedures for Missing-Data Problems. *Biometrika Trust*, **70**, 613-624. http://dx.doi.org/10.1093/biomet/70.3.613

[7] Hocking, R.R. and Smith, W.B. (2000) Estimation of Parameters in the Multivariate Normal Distribution with Missing Observations. *Journal of the American Statistical Association*, No. 63, 159-173.

[8] Anderson, T.W. and Olkin, I. (1985) Maximum-Likelihood Estimation of the Parameters of a Multivariate Normal Distribution. *Linear Algebra and Its Applications*, **70**, 147-171. http://dx.doi.org/10.1016/0024-3795(85)90049-7

[9] Fisher, R.A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals Eugenics*, **7**, 179-188. http://dx.doi.org/10.1111/j.1469-1809.1936.tb02137.x