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

Phil D. Young1, Dean M. Young2, Songthip T. Ounpraseuth3
1Department of Information Systems, Baylor University, Waco, TX, USA
2Department of Statistical Science, Baylor University, Waco, TX, USA
3Department 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

Copyright © 2016 by authors and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).

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 \( x \sim N_p(\mu, \Sigma) \) into one of two distinct
multivariate normally distributed populations $\Pi_i : N_p(\mu_i, \Sigma), \ i = 1, 2$, when monotone missing training data are present, where $\mu_i$ and $\Sigma$ 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 $x_i$ when, if $x_{ij}$ is missing, then $x_{ik}$ 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 EERs. 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 $\mathbb{R}^{n \times n}$ to represent the matrix space of all $n \times n$ matrices over the real field $\mathbb{R}$. Also, we let the symbol $\mathbb{R}^+_{n \times n}$ represent the cone of all $n \times n$ positive definite matrices in $\mathbb{R}^{n \times n}$. Moreover, $A' \in \mathbb{R}^{n \times n}$ represents the transpose of $A \in \mathbb{R}^{n \times n}$.

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 EERs 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

$$
\begin{bmatrix}
Y_{1i} & Y_{2i} \\
Z_i & 
\end{bmatrix},
$$

(1)

where

$$
U_i = \left[ Y_{1i}' : Z_i' \right] \in \mathbb{R}_{k \times n_i}
$$

(2)

denotes the $n_i$ complete-observation submatrix, and $Y_{2i} \in \mathbb{R}_{k \times (n_i - n_j)}$ 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 $u_j = \left[ y_{1j}' : z_j' \right]$, where $y_{1j} \in \mathbb{R}_{k \times 1}$ and $z_j \in \mathbb{R}_{(p-k) \times 1}$ such that

$$
u_j \sim N_p(\mu_j, \Sigma) = N_p\left(\begin{bmatrix} \mu_{j1} \\ \mu_{j2} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),
$$

(3)

where $\mu_{j1} \in \mathbb{R}_{k \times 1}$, $\mu_{j2} \in \mathbb{R}_{(p-k) \times 1}$, $\Sigma_{11} \in \mathbb{R}^{k \times k}$, $\Sigma_{12} \in \mathbb{R}^{k \times (p-k)}$, and $\Sigma_{22} \in \mathbb{R}^{(p-k) \times (p-k)}$ with $i = 1, 2, j = 1, 2, \ldots, n_j$. Also, random samples of sizes $N_i - n_i$ are taken from distributions of the form $N_p(\mu_{ji}, \Sigma_{ij})$, where $\mu_{ji} \in \mathbb{R}_{k \times 1}$ and $\Sigma_{ij} \in \mathbb{R}^{k \times k}$.

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 \(U_i\), \(i = 1, 2\), given in (2), which is

\[ W_u = (\overline{u}_i - \overline{u}_j) S_u^{-1} \left[ u - \frac{1}{2}(\overline{u}_i + \overline{u}_j) \right], \]

where \(\overline{u}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} u_{ij}\) and

\[ S_u = \frac{1}{n_1 + n_2 - 2} \left( \sum_{i=1}^{n_1} \sum_{j=1}^{n_i} (u_{ij} - \overline{u}_i)(u_{ij} - \overline{u}_j) \right), \]

are the complete-data sample mean and complete-data sample covariance matrix, respectively. They also use Anderson’s \(LDF\) for the data

\[ [Y_{i1} : Y_{i2}], \quad i = 1, 2, \] with \(k\) features and \(N_i + N_2\) training observations, which is

\[ W_f = (\overline{y}_i - \overline{y}_j) S_f^{-1} \left[ y - \frac{1}{2}(\overline{y}_i + \overline{y}_j) \right] \]

with

\[ \overline{y}_i = \frac{1}{N_i} \left[ n_i \overline{y}_{i1} + (N_i - n_i) \overline{y}_{i2} \right], \]

where

\[ \overline{y}_{i1} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \]

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

\[ \overline{y}_{i2} = \frac{1}{N_i - n_i} \sum_{j=n_i+1}^{N_i} y_{ij} \]

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

\[ S_f = \frac{1}{N_1 + N_2 - 2} \left( \sum_{i=1}^{N_1} \sum_{j=1}^{N_i} \sum_{i'=1}^{N_1} \sum_{j'=1}^{N_i'} (y_{i'j'} - \overline{y}_{i'})(y_{i'j'} - \overline{y}_{i'}) \right) \]

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

\[ W_c = c W_u + (1-c) W_f, \]

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

\[ W_c \geq 0 \]

and into \(\Pi_j\), otherwise. The conditional error rate (CER) for classifying an unlabeled vector \(x\) from \(\Pi_i\) into \(\Pi_j\) using (8) is

\[ CER_{ij}(W_c) = P\left( (-1)^{i-j} W_c < 0 \mid \overline{u}_i, \overline{u}_j, S_u, \overline{y}_i, \overline{y}_j, S_f; u, y \in \Pi_i \right) \]

\[ = \Phi \left( \frac{(-1)^{i-j} \mu_i \mu_j + (-1)^{i-j} f}{\sqrt{h_i} \Sigma_h} \right), \]

\(i, j = 1, 2, \quad i \neq j\), where

\[ f = cb + (1-c)e \]
with
\[ b = -\frac{1}{2}(\overline{u}_1 - \overline{u}_2)' S_u^{-1} (\overline{u}_1 + \overline{u}_2) \text{ and } e = -\frac{1}{2}(\overline{y}_1 - \overline{y}_2)' S_y^{-1} (\overline{y}_1 + \overline{y}_2). \]

Also,
\[ h = [r' : b']', \]
where \( r \in \mathbb{R}_{k \times 1} \) and \( b \in \mathbb{R}_{(p-k) \times 1} \), \( r = ca_1 + (1-c)d \), \( b = ca_2 \), \( d = S_y^{-1} (\overline{y}_1 - \overline{y}_2) \), \( a = S_u^{-1} (\overline{u}_1 - \overline{u}_2) = [a_1' : a_2'] \), with \( a_1 \in \mathbb{R}_{k \times 1} \), \( a_2 \in \mathbb{R}_{(p-k) \times 1} \), and \( \overline{y}_j \) defined in (5). Thus, using (9) and assuming equal \textit{a priori} probabilities, the CER for (8) is
\[ \text{CER}(W_c) = \frac{1}{2} [CER_{12}(W_c) + CER_{22}(W_c)]. \]

If
\[ \hat{\theta} = [\overline{y}_1 : \overline{y}_2 : S_y : S_u : \overline{u}_1 : \overline{u}_2], \]
then, for (8), the EER of misclassifying an unlabeled observation vector \( x \) from \( \Pi_i \) into \( \Pi_j \) is
\[ \text{EER}(W_c) = E_{\theta} \left[ \Phi \left( \frac{(-1)^{2-i} h' \mu + (-1)^{2-i} f}{h \Sigma h} \right) \right]. \]

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 CERs, at least one of the sample Mahalanobis distances
\[ D_u = (\overline{u}_1 - \overline{u}_2)' S_u^{-1} (\overline{u}_1 - \overline{u}_2), \]
will be large. While \( n_i \) and \( D_u \) determine the performance of \( W_u \), the quantities \( N_f \) and \( D_f \) dictate the performance of \( W_y \). 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 \( D_u > D_f \) is that the information in the data-matrix component \( Z_i \), \( i = 1, 2 \), in (1) contributes largely to the discriminatory information. Hence, [1] uses
\[ c' = \frac{1 + \frac{1}{n_1}}{n_1 + n_2} D_u^{-1} + \frac{1 + \frac{1}{N_1}}{N_1 + N_2} D_f^{-1} \]
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 \textit{MLE} method for estimating parameters in a multivariate normal distribution with \textit{BMM} data. The estimator of \( \Sigma \) in the [7] \textit{MLEs} classifier is a pooled estimator of the two individual \textit{MLEs} of \( \Sigma \).

Below, we state the \textit{MLEs} for two multivariate normal distributions having unequal means and a common covariance matrix with identical \textit{BMM}-data patterns in both training samples.

**Theorem.** Let \( \Pi_i \) be modeled by the multivariate normal densities \( N_i(\mu, \Sigma) \) for \( i = 1, 2 \), with
\[ \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \]
(13)
and

\[ \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \] (14)

Also, let

\[ A_{1,N_i,j} = \sum_{j=1}^{N_i} (y_j - \bar{y}_i)(y_j - \bar{y}_i)^\prime, \]

\[ A_{1,N_i,i} = \sum_{j=1}^{N_i} (y_j - \bar{y}_i)(y_j - \bar{y}_i)^\prime, \]

\[ A_{12,N_i,j} = \sum_{j=1}^{N_i} (y_j - \bar{y}_i)(z_j - \bar{z}_i)^\prime, \]

and

\[ A_{22,N_i,j} = \sum_{j=1}^{N_i} (z_j - \bar{z}_i)(z_j - \bar{z}_i)^\prime, \]

where \( y_j \in [Y_1 : Y_2] \), and \( z_j \in Z \), with \( Y_1, Y_2 \) and \( Z \) given in (1). Then, on the basis of two-step monotone training samples from populations \( \Pi_i : N_i (\mu_i, \Sigma), i = 1, 2 \), the MLEs of (13) and (14) are

\[ \hat{\mu} = \left[ \hat{\mu}_1, \hat{\mu}_2 \right] \quad \text{and} \quad \hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix}, \] (15)

respectively, where

\[ \hat{\Sigma}_{11} = \frac{\sum_{j=1}^{N_i} A_{1,N_i,j}}{\sum_{j=1}^{N_i} N_i}, \]

\[ \hat{\Sigma}_{12} = \frac{1}{\sum_{j=1}^{N_i} N_i} \left[ \sum_{j=1}^{N_i} A_{1,N_i,j} \right] \left[ \sum_{j=1}^{N_i} A_{1,i,j} \right]^{-1} \left[ \sum_{j=1}^{N_i} A_{12,N_i,j} \right], \] (16)

and

\[ \hat{\Sigma}_{22} = \frac{1}{\sum_{j=1}^{N_i} N_i} \sum_{j=1}^{N_i} A_{22,N_i,j} + \frac{1}{\sum_{j=1}^{N_i} N_i} \left[ \sum_{j=1}^{N_i} A_{21,N_i,j} \right] \left[ \sum_{j=1}^{N_i} A_{11,N_i,j} \right]^{-1} \left[ \sum_{j=1}^{N_i} A_{12,N_i,j} \right], \] (17)

with \( \hat{\mu}_i = \bar{y}_i \), where \( \bar{y}_i \) is defined in (5),

\[ \hat{\mu}_2 = z_i - \hat{\Sigma}_{12} \hat{\Sigma}_{22}^{-1} (\bar{y}_1 \bar{y}_2), \]

\[ \bar{z}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} z_j, \]

and

\[ \sum_{j=1}^{N_i} A_{21,N_i,j} = \sum_{j=1}^{N_i} A_{22,N_i,j} - \left[ \sum_{j=1}^{N_i} A_{21,N_i,j} \right] \left[ \sum_{j=1}^{N_i} A_{11,N_i,j} \right]^{-1} \left[ \sum_{j=1}^{N_i} A_{12,N_i,j} \right]. \]
Proof: A proof is alluded to in [8].

The MLES classification statistic is

\( W_{MLE} = (\hat{\mu}_2 - \hat{\mu}_1) \mathbf{\Sigma}^{-1} \left[ x - \frac{1}{2} (\hat{\mu}_2 + \hat{\mu}_1) \right] \) \tag{18}

where \( \hat{\mu}_1, \hat{\mu}_2, \) and \( \hat{\Sigma} \) are the MLEs defined in (15), and \( x \in \mathbb{R}^p \) is an unlabeled observation vector belonging to either \( \Pi_1 \) or \( \Pi_2 \). We classify the unlabeled observation vector \( x \in \mathbb{R}^p \) into \( \Pi_1 \) if

\( W_{MLE} \leq 0 \) \tag{19}

and into \( \Pi_2 \), otherwise. Given that \( x \in \Pi_1 \), conditioning on \( \hat{\mu}_j, i, j = 1, 2 \), and \( \hat{\Sigma} \), and using the fact that

\( \hat{\delta}^\prime \hat{\Sigma}^{-1} (x - \mu) \sim N(0, \hat{\delta}^\prime \hat{\Sigma}^{-1} \hat{\Sigma}^{-1} \hat{\delta}) \),

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

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

where

\[ w_i = \left\{ \frac{\hat{\delta}^\prime \hat{\Sigma}^{-1} \hat{\Sigma}^{-1} \hat{\delta}}{\sqrt{2}} \left\{ \hat{\delta}^\prime \hat{\Sigma}^{-1} \left[ \frac{1}{2} (\hat{\mu}_2 + \hat{\mu}_1) - \hat{\mu}_i \right] \right\} \right\} \quad (20) \]

\( i = 1, 2 \). Similarly, given \( x \in \Pi_2 \),

\( CER_{21} \left( \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma} \right) = P \left[ W_{MLE} \leq 0 | \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma} ; x \in \Pi_2 \right] = \Phi(w_i) \),

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

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

Hence, the overall expected error rate is

\( EER \left( \hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma} \right) = \frac{1}{2} \left[ 1 - F_{\hat{\theta}}(\Phi(w_1)) + F_{\hat{\theta}}(\Phi(w_2)) \right] \).

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, \( \Pi_i : N_p (\mu, \Sigma) \), \( i = 1, 2 \), 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 \( \Pi_i, i = 1, 2 \), and

\( \Sigma = \rho J + (1 - \rho) I \)

is the intraclass covariance matrix where \( \rho \) 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 CERs were calculated for each training-sample set. Then, the differences between the CERs for the classifiers were averaged over the 10,000 CER differences for the two classifiers for each parameter configuration involving \( N_i \), \( p \), \( r \), \( \Sigma \), \( \mu \), and POMD for the \( r \) features with monotone missing data, where \( i = 1, 2 \). Thus, the \( \overline{\text{EERD}} \) for the C-H and MLES classifiers is

\[
\overline{\text{EERD}} = \frac{1}{K} \sum_{j=1}^{K} \left[ \text{CER}_j(W) - \text{CER}_j(\hat{\mu}, \hat{\Sigma}) \right],
\]

where \( \text{CER}(W) \) is defined in (12), \( \text{CER}(\hat{\mu}, \hat{\mu}, \hat{\Sigma}) \) 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 \( j \in \{1, 2, \ldots, K\} \). We display the results of our two Monte Carlo simulations by graphing \( \text{EERD} \) against \( p = 0, 0.1, 0.3, 0.5, 0.7, 0.9 \) for various configurations of \( p, r \), \( N_i \), \( \mu \), \( \Sigma \), 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 \( \overline{\text{EERD}} \) 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 \( \mu_i \in \mathbb{R}_{p = 1} \) such that

\[
\mu_i = \left[ 0, 0, \ldots, 0 \right] \quad (22)
\]

and \( \mu_2 \in \mathbb{R}_{p = 1} \) such that

\[
\mu_2 = \left[ d_j, 0, 0, \ldots, 0, d_j, 0, \ldots, 0 \right], \quad (23)
\]

with \( d_1 = 0.5 \) and \( d_2 = 3 \) to assess \( \overline{\text{EERD}} \) for both small and large between-class separation. These values for \( \mu_i \), \( i = 1, 2 \), 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 \( \overline{\text{EERD}} \) is plotted against \( \rho \) 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 \( \text{EERD} \) 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 \( \overline{\text{EERD}} \) is consistently positive for the values of \( p, r, N_i, \rho, d_j, \) and POMD examined here. Moreover, the figures indicate that the primary parameters that influence the dominance of the MLES classifier are \( \rho \) 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 \( \rho > 0.1 \), \( \text{EERD} \) increased as \( \rho \) increased for all \( p, r, N_i, \mu, \mu \), and POMD considered here. The most noteworthy increase in the \( \overline{\text{EERD}} \) was for \( 0.7 \leq \rho \leq 0.9 \) when \( d_i = 0.5 \), where \( \overline{\text{EERD}} \) 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 \( \text{EERD} \) for these configurations. Additionally, we noted that \( \text{EERD} \approx 0.20 \) when \( \rho = 0.9, d_i = 0.5 \), and other parameters are allowed to vary.

The MLES classifier especially outperformed the C-H classifier when \( d_i = 0.5 \) for \( \rho > 0 \), as compared to when \( d_2 = 3 \). The smaller values of \( \text{EERD} \) 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 |
Mahalanobis distance when $d_2 = 3$ and $\rho = 0.1$, the EERs for both classifiers are small, thus yielding a smaller $\overline{EERD}$.

As we used a large number of simulation iterations, we obtained $\max_\xi \{\text{s.e.}(\overline{EERD})\} < 0.003$, where $\xi$ is the grid of parameter vectors considered in the simulation. Thus, the relatively small estimated standard errors also support our claim that $\overline{EERD} > 0$ 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 $\overline{EERD}$. Regardless of the combination of parameter values considered here, the $MLES$ classifier dominates the $C-H$ classifier in terms of $\overline{EERD}$.

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 $\overline{EERD}$ in the [1] simulations are not sufficiently small enough to conclude a difference in the ERRs 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 $\mu_2$ in (23), where we chose 0.80 of the elements of $\mu_2$ 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.
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 EERs of the C-H and MLE 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 EER estimator for the C-H classifier, $EER_{\text{boot}(C-H)}$. Let $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\Sigma}$ be the MLEs of $\mu_1$, $\mu_2$, and $\Sigma$, respectively, defined in Theorem 1. Also, let $\hat{\mu}_1^*$, $\hat{\mu}_2^*$, and $\hat{\Sigma}^*$ be the bootstrap estimates of $\hat{\mu}_1$, $\hat{\mu}_2$, and $\hat{\Sigma}$, respectively, calculated using the parametric bootstrap training-sample data

$$
\begin{bmatrix}
Y_{i1}^* & Y_{i2}^* \\
Z_i^*
\end{bmatrix}
$$

(24)

that is generated from $N_p\left(\hat{\mu}_i, \hat{\Sigma}\right)$, $i = 1, 2$. Then, conditioning on $\hat{\mu}_1^*$, $i = 1, 2$, and $\hat{\Sigma}^*$, the bootstrap CERs for the C-H classifier are

$$
CER_{\text{boot}}^i(W_j^*) = \Phi\left(\frac{(-1)^{\delta_{ij}^*} h^* \hat{\mu}_j^* + (-1)^{\delta_{ij}^*} f^*}{\sqrt{h^* \hat{\Sigma}^* h^*}}\right)
$$

for $i, j = 1, 2, i \neq j$, where $W_j^*$, $h^*$, and $f^*$ are similar in definition to $W_i^*$, $h$, 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
\[ CER^*(W') = \frac{1}{2} \left[ CER_{12}^*(W') + CER_{21}^*(W') \right]. \]  

(25)

Also, conditioning on \( \hat{\mu}', i = 1, 2, \) and \( \hat{\Sigma}', \) the bootstrap CERs for the MLES classifier are

\[ CER_j^*(\hat{\mu}', \hat{\mu}', \hat{\Sigma}') = P\left[ (-1)^{2-j} W_{MLE}^* > 0 \mid \hat{\mu}', \hat{\mu}', \hat{\Sigma}'; x \in \Pi \right], \]

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

\[ CER_{12}^*(\hat{\mu}', \hat{\mu}', \hat{\Sigma}') = 1 - \Phi\left( w_i' \right), \]

and given \( x \in \Pi_2, \)

\[ CER_{21}^*(\hat{\mu}', \hat{\mu}', \hat{\Sigma}') = \Phi\left( w_2' \right), \]

where

\[ w_i' = \left[ \delta'^* \hat{\Sigma}^{-1} \Sigma^{-1} \delta' \right]^{-1/2} \left[ \delta'^* \hat{\Sigma}^{-1} \left( \frac{1}{2} \hat{\mu}' - \bar{\mu} \right) \right], i = 1, 2. \]

Thus, assuming equal a priori probabilities of belonging to \( \Pi_i, \) \( i = 1, 2, \) for an unlabeled observation, we have

\[ CER^*(\hat{\mu}', \hat{\mu}', \hat{\Sigma}') = \frac{1}{2} \left[ 1 - \Phi\left( w_1' \right) + \Phi\left( w_2' \right) \right]. \]  

(26)

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

\[ \tilde{EERD}_{\text{boot}} = \frac{1}{K} \sum_{j=1}^{K} \left( CER_{\text{boot}}(C-H) - \tilde{CER}_{\text{boot}}(MLES) \right), \]

(27)

where \( j \) denotes the \( j^{th} \) simulated training-data set for \( j \in \{1, 2, \ldots, K\}. \) 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 (\( \Pi_1 \)) and the Failure Group for students who do not complete their master’s degrees (\( \Pi_2 \)). 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

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

(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 \( EERD_{\text{boot}} \), given in (27), using 10,000 bootstrap simulation iterations with \( p = 5, \ r = 1, \ N_i = 20, \)
Table 2. UTA Admissions office.

<table>
<thead>
<tr>
<th>$\Pi_1$ : Success</th>
<th>$\Pi_1$ : Failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$x_2$</td>
</tr>
<tr>
<td>2.97</td>
<td>420</td>
</tr>
<tr>
<td>3.80</td>
<td>330</td>
</tr>
<tr>
<td>2.50</td>
<td>270</td>
</tr>
<tr>
<td>2.50</td>
<td>400</td>
</tr>
<tr>
<td>3.30</td>
<td>280</td>
</tr>
<tr>
<td>2.60</td>
<td>310</td>
</tr>
<tr>
<td>2.70</td>
<td>360</td>
</tr>
<tr>
<td>3.10</td>
<td>220</td>
</tr>
<tr>
<td>2.60</td>
<td>350</td>
</tr>
<tr>
<td>3.20</td>
<td>360</td>
</tr>
<tr>
<td>3.65</td>
<td>440</td>
</tr>
<tr>
<td>3.56</td>
<td>640</td>
</tr>
<tr>
<td>3.00</td>
<td>480</td>
</tr>
<tr>
<td>3.18</td>
<td>550</td>
</tr>
<tr>
<td>3.84</td>
<td>450</td>
</tr>
<tr>
<td>3.18</td>
<td>410</td>
</tr>
<tr>
<td>3.43</td>
<td>460</td>
</tr>
<tr>
<td>3.52</td>
<td>580</td>
</tr>
<tr>
<td>3.09</td>
<td>450</td>
</tr>
<tr>
<td>3.70</td>
<td>420</td>
</tr>
</tbody>
</table>

and $\eta_i = 10$ for $i = 1, 2$. Additionally, the parametric bootstrap multivariate normal distribution parameters, which are the MLEs for the multivariate normal population parameters given in Theorem 1, are

$$\hat{\mu}_1 = \begin{bmatrix} 3.171, 409, 644, 526.25, 577.01 \end{bmatrix}$$

and

$$\hat{\mu}_2 = \begin{bmatrix} 3.087, 430, 649, 519.75, 562.66 \end{bmatrix}$$

for the means of $\Pi_1$ and $\Pi_2$, respectively, with common covariance matrix


Subsequently, we obtained $\overline{EERD}_{\text{Boot}} = -0.027$ with $\text{se}(\overline{EERD})_{\text{Boot}} = 0.001$, 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 $MLES$ 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 ($\Pi_1$), Iris-versicolor ($\Pi_2$), and Iris-virginica ($\Pi_3$). We have used a subset of the original Iris data set by taking only the first 20 observations from $\Pi_1$ and $\Pi_2$ and omitting the Iris-virginica group ($\Pi_3$). 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

$$
\hat{C}_{\text{iris}} = \begin{bmatrix}
1 & 0.716 & 0.708 & 0.549 \\
0.716 & 1 & 0.473 & 0.651 \\
0.708 & 0.473 & 1 & 0.677 \\
0.549 & 0.651 & 0.677 & 1
\end{bmatrix}
$$

#### Table 3. Partial iris data.

<table>
<thead>
<tr>
<th></th>
<th>$\Pi_1$: Setosa</th>
<th>$\Pi_2$: Versicolor</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>5.1</td>
<td>4.9</td>
</tr>
<tr>
<td>$x_2$</td>
<td>3.5</td>
<td>3.0</td>
</tr>
<tr>
<td>$x_3$</td>
<td>1.4</td>
<td>1.4</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>$x_1$</td>
<td>7.0</td>
<td>6.4</td>
</tr>
<tr>
<td>$x_2$</td>
<td>3.2</td>
<td>2.8</td>
</tr>
<tr>
<td>$x_3$</td>
<td>4.7</td>
<td>4.5</td>
</tr>
<tr>
<td>$x_4$</td>
<td>1.4</td>
<td>1.5</td>
</tr>
</tbody>
</table>
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 $\overline{EERD}_{\text{boot}}$. 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 $\Pi_1$ and $\Pi_2$ are

$$\hat{\mu}_1 = [5.035, 3.48, 1.435, 0.235]$$

and

$$\hat{\mu}_2 = [5.975, 2.76, 4.255, 1.325],$$

respectively, with common covariance matrix

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

For the parametric bootstrap estimate for $\overline{EERD}$ corresponding to the C-H and MLES classifiers applied to the subset of the Iris data set, we obtained $\overline{EERD}_{\text{boot}} = 0.11$ with $\overline{s.e.}(\overline{EERD}_{\text{boot}}) = 0.001$, which indicated that $EERD_{\text{MLE}} \ll EERD_{\text{(C-H)}}$. 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, $\rho$, 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 $\rho$ 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 $\overline{EERD}_{\text{boot}}$ 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


