ABSTRACT

Topics in Dimension Reduction and Missing Data in Statistical Discrimination

Phil D. Young, Ph.D.

Chairperson: Jack D. Tubbs, Ph.D.

This dissertation is comprised of four chapters. In the first chapter, we define the concept of linear dimension reduction, review some popular linear dimension reduction procedures, discuss background research that we use in chapters two and three, and give a brief outline of the dissertation contents.

In chapter two, we derive a linear dimension reduction (LDR) procedure for statistical discriminant analysis for multiple multivariate skew-normal populations. First, we define the multivariate skew-normal distribution and give several applications of its use. We also provide marginal and conditional properties of the MSN random vector. Then, we state and prove several lemmas used in a series of theorems that present our LDR procedure for the multivariate skew-normal populations using parameter configurations. Lastly, we illustrate our LDR method for multiple multivariate skew-normal distributions with three examples.

In the third chapter, we define and rigorously prove the existence of the multivariate singular skew-normal ($MSSN$) distribution. Next, we state and prove distributional properties for linear combinations, marginal, and conditional random variables from a $MSSN$ distribution. Then, we state and prove several lemmas used in deriving our LDR transformation for the multiple $MSSN$ distributions with assorted parameter combinations. We then state and prove several theorems con-

cerning the formulation of our *LDR* technique. Finally, we illustrate the effectiveness of our *LDR* technique for multiple multivariate singular skew-normal classes with two examples.

In chapter four, we compare two statistical linear discrimination procedures when monotone missing training data exists in the training data sets from two different multivariate normally distributed populations with unequal means but equal covariance matrices. We derive the maximum likelihood estimators (*MLE*s) for the partitioned population means and the common covariance matrix in an appendix. Additionally, we contrast two classifiers: a linear combination discriminant function derived from Chung and Han (*C-H*) (2000) and a linear classifier based on the *MLE* of two multivariate normal training samples with identical monotone missing training-data in one or more features. We then perform two Monte Carlo simulations with various parameter configurations to compare the effectiveness of the *MLE* and *C-H* classifiers as the correlation between features for the population covariance matrix increases. Moreover, we compare the two competing classifiers using parametric bootstrap estimated expected error rates for a subset of the well-known Iris data.

Topics in Dimension Reduction and Missing Data in Statistical Discrimination

by

Phil D. Young, B.S., M.S.

A Dissertation

Approved by the Department of Statistical Science

_____

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

_____

Jack D. Tubbs, Ph.D., Chairperson

_____

James D. Stamey, Ph.D.

_____

Tom L. Bratcher, Ph.D.

_____

David J. Ryden, Ph.D.

_____

Dean M. Young, Ph.D.

Accepted by the Graduate School
December 2009

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank God, who has blessed both my family and me in so many ways. His guidance, unconditional love, and abounding compassion have kept me going during the tough times.

Next, I would like to thank my family for every bit of love and support that they have given me throughout the years. Mom, Dad, and Alissa, I love you so much and want you to know that your constant support has meant the world to me. Mom, thank you for everything you have done for me. You have done so much. To my sister, Alissa, I also want to thank you for sticking by me throughout the years. I have really enjoyed getting to know you and am happy that we have grown closer together in recent years.

To the entire faculty from the Department of Statistical Sciences, I would like to thank you very much. To Dr. Hill, Dr. Stamey, Dr. Bratcher, Dr. Seaman, Dr. Tubbs, Dr. Harvill, Dr. Johnson, and Dr. Maddox, I would like to say thank you for taking the time to teach me. Whether it was inside or outside of the classroom, you were always willing to answer questions and offer your knowledge.

To other professors outside of this department, I would like to thank Dr. Brian Raines, Dr. David Ryden, and Wes Evans. Thank you for showing me that I had the potential to be successful in the field of mathematics, and thank you for your inspiration. If you had not taken somewhat of an interest in me, I would have given up a long time ago.

To my fellow students in the department, thank you for all your help and support. Thank you for the laughs, the help on homework, and, especially, the help on the computer. I would especially like to thank Johnny Seaman and Daniel Beavers for your endless supply of laughter and good humor when it was very much

in demand. Also, thank you to Stephanie Powers, Jo Wick, and Brandi Greer, who have always been willing to help me whenever help was needed. The people I have worked with have been smart, talented, and considerate, which has helped me considerably in being the best that I can be.

To my best friends outside of the department, namely Derek Barnett, Joe Khurana, Cody Smith, and D.B. Briscoe, I would like to thank you for friendship throughout the years. Though your help has been unrelated to statistics, your advice, encouragement, and moral support have been solid and uplifting.

Finally, I would like to thank my father, Dean Young. Dad, you are a great father, teacher, and friend. I want to thank you from the bottom of my heart for sticking with me and helping me through this long and arduous journey. Words cannot express how much I care about you or how much I appreciate all the hard work you have put into working with me. I know that without a father like you, I would be completely lost. Thank you once again, and I look forward to working with you.

DEDICATION

To my family

CHAPTER ONE

A Review of Dimension Reduction and Various Dimension Reduction Methods for Supervised Classification

1.1 Introduction

An escalation in data collection procedures and an expansion in storage capabilities have made data dimension reduction necessary in recent years. Dimension reduction is required to transform high-dimensional data into significantly smaller dimensional data without incurring a significant loss of information.

We formulate the dimension reduction problem as follows. Suppose we obtain the p-dimensional data matrix

$$\mathbf{X}_{p\times N} \equiv [\mathbf{x}_{11}, \mathbf{x}_{12}, ..., \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \mathbf{x}_{22}, ...\mathbf{x}_{2n_2}, ..., \mathbf{x}_{m1}, \mathbf{x}_{m2}, ..., \mathbf{x}_{mn_m}],$$

containing training data sampled from m distinct populations where $\mathbf{x}_{ij} \in \mathbb{R}_{p\times 1}$ for $i \in \{1, 2, ..., m\}$ and $j \in \{1, 2, ..., n_i\}$, with $N = \sum_{i=1}^{m} n_i$ observations. Here, $\mathbf{x}_{ij}$ denotes the j^{th} observation for the i^{th} class. A dimension reduction technique transforms the data matrix $\mathbf{X}$ into a q-dimensional data matrix

$$\mathbf{Y}_{q\times N} \equiv [\mathbf{y}_{11}, \mathbf{y}_{12}, ..., \mathbf{y}_{1n_1}, \mathbf{y}_{21}, \mathbf{y}_{22}, ...\mathbf{y}_{2n_2}, ..., \mathbf{y}_{m1}, \mathbf{y}_{m2}, ..., \mathbf{y}_{mn_m}]$$

with $\mathbf{y}_{ij} \in \mathbb{R}_{q\times 1}$, where $q \ll p$, while retaining the geometric characteristics of the data in the process. That is, we seek a transformation $T : \mathbb{R}_{p\times 1} \to \mathbb{R}_{q\times 1}$ such that $T(\mathbf{x}_{ij}) = \mathbf{y}_{ij}$ for $i \in \{1, 2, ..., m\}$ and $j \in \{1, 2, ..., n_i\}$.

Dimension reduction can be categorized into two basic forms: linear and nonlinear. Nonlinear dimension reduction ($NLDR$) techniques are advantageous for nonlinear data. In $NLDR$, we wish to retain the discriminatory information of high-dimensional data by reducing the original data to a lower-dimensional form when our data points lie on a nonlinear manifold. Some popular $NLDR$ techniques

include multidimensional scaling, isomaps, kernel principal components analysis, diffusion maps, locally linear embedding, and Laplacian eigenmaps. However, in this dissertation, we focus only on linear dimension reduction LDR.

We can partition LDR into two main approaches: LDR and feature subset selection. Feature subset selection selects a subset of original variables that are relevant for constructing robust learning models. The feature subset selection process helps the researcher better understand the data by indicating which features are important. Once chosen, the dimensions retained by feature subset selection are directly interpretable. On the other hand, linear feature selection is the process of mapping elements of high-dimensional space into a lower-dimensional subspace using a linear transformation. In LDR, a high-dimensional observation $\mathbf{x}_{ij} \in \mathbb{R}_{p\times 1}$ is transformed to a small dimension by the linear transformation $\mathbf{y}_{ij} = \mathbf{K}\mathbf{x}_{ij}$, where the matrix $\mathbf{K} \in \mathbb{R}_{q\times p}$ represents a linear projection from the p-dimensional feature space to the lower q-dimensional transformed space. If the dimension reduction performs well, we expect that most of the relevant classification information will be contained in the reduced data.

An additional application of feature selection is data visualization. A researcher may have an extraordinary ability to recognize systematic patterns in a dataset but generally be unable to adequately interpret a dataset if the data dimension is greater than three. To permit the visualization of high-dimensional data, we frequently wish to choose two or three of the most informative transformed features in the dataset and plot them so that we can envision the data relationships in a reduced dimension.

We have several reasons to perform dimension reduction. For instance, Cunningham (2007) has claimed that implementing dimension reduction can reveal new knowledge about a dataset. Furthermore, data vectors may contain many highly correlated variables, and the pertinent data might be explicable with only a few

variables. Using dimension reduction, we can determine the most important features to explain the essential phenomenon of the data while simultaneously eliminating the redundant information. Dimension reduction is often applied even when one is dealing with extremely high-dimensional and computationally advanced models. Pavlenko (2003) has stated that another motivation for dimension reduction is when "expensive" measurement occurs. Here, the term "expensive" implies "costly" with respect to time, money, or computational speed. Pavlenko further states that "the omission of certain features or sets of features, while naturally destroying the possible optimality of standard discriminant analysis, will not seriously affect the error probability or any other criterion of interest."

Researchers implement dimension reduction for supervised classification in a wide variety of fields, including engineering, astronomy, biology, face and voice recognition, remote sensing, economics, consumer transactions, and microarray data. Hamsici and Martinez (2008) have discussed methods in which dimension reduction can be administered in numerous disciplines. For example, for $p > 2$, psychologists and anthropologists may desire a two-dimensional visualization from a collection of n p-dimensional sample observations belonging to m classes in order to draw inferences about the different groups. Also, one goal of medical diagnosis is to determine a specific combination of factors that describe a set of m subclasses of a disease. Specifically, different treatments are required for each stage of cancer, and, therefore, one needs to extricate the primary factors used for classifying stages of cancer to ensure the proper treatment. With the utilization of dimension reduction, we can determine the ideal combination of features by reducing the original p-dimensional data to q components most indicative of the disease level. Also, in computer vision and speech analysis, we wish to find the smallest-dimensional subspace where the class distributions can be efficiently separated. This q-dimensional subspace is preferable to the original p-dimensional data because lower-dimensional representa-

tions of multivariate populations for learning algorithms generally perform better, faster, and more efficiently.

Another motivation for dimension reduction is known as the curse of dimensionality. The term "curse of dimensionality" was coined by Richard Bellman (1961), and, according to Donoho (2000), refers to Bellman's assertion that "if our goal is to optimize a function over a continuous product domain of a few dozen variables by exhaustively searching a discrete space defined by a crude discretization, we could easily be faced with the problem of making tens of trillions of evaluations of the function." According to Chen (2007), the curse of dimensionality alludes to the condition in which we lack the essential assumptions to simplify our model so that the sample size needed to accurately estimate our multivariate dataset is exponentially amplified as the dimension increases. The curse of dimensionality causes an extremely low rate of convergence when one attempts to approximate a statistical function in high dimensions. Carreira-Perpinan (1997) affirms that the curse of dimensionality often induces the existence of correlations in large-dimensional feature vectors. Hence, even with a large sample size, estimating the density of the data provides a questionably large optimal integrated squared mean. However, one can often circumvent this dimensionality curse to estimate a statistical function with the assistance of a dimension reduction technique.

1.2 Some Common LDR Methods for Statistical Supervised Classification

The first known LDR method for statistical classification was derived by R. A. Fisher (1936), who originated the linear discriminant function (LDF) for two classes. Let $\mathbf{w} \in \mathbb{R}_{p\times 1}$. The optimization criterion for Fisher's LDF is as follows:

$$\max_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}'\mathbf{S}_b\mathbf{w}}{\mathbf{w}'\mathbf{S}_w\mathbf{w}},$$

where

$$\mathbf{S}_b = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'$$

is the between-class scatter matrix, and

$$\mathbf{S}_w = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{x}_j - \bar{\mathbf{x}}_i)(\mathbf{x}_j - \bar{\mathbf{x}}_i)'$$

is the within-class scatter matrix. Also, $\mathbf{w}$ is defined to be the generalized eigenvector,

$$\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \tag{1.1}$$

is the i^{th} class sample mean, and $\mathbf{x}_{ij}$ is the j^{th} observation vector for group i, where $i = 1, 2$, and $j = 1, 2, ..., n_i$.

In 1948, C. R. Rao generalized Fisher's LDF to multiple classes in his doctoral dissertation, written under Fisher, and published his results in Rao (1948). This classification and dimension reduction method is known as linear discriminant analysis (LDA). The goal of LDA is to produce a linear transformation that maximizes the ratio of the average between-class scatter matrix relative to the average within-class scatter matrix. Essentially, we wish to determine $\mathbf{W} \in \mathbb{R}_{q \times p}$ such that the Fisher criterion,

$$\max_{\mathbf{W} \neq \mathbf{0}} \frac{\mathbf{W}'\mathbf{S}_B\mathbf{W}}{\mathbf{W}'\mathbf{S}_W\mathbf{W}},$$

is maximized, where

$$\mathbf{S}_W = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'$$

and

$$\mathbf{S}_B = \sum_{i=1}^{m} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$$

denote the within-class and between-class scatter matrices, respectively. Also,

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{x}_{ij}$$

denotes the overall sample mean, $\bar{\mathbf{x}}_i$ is defined in (1.1), $N = \sum_{i=1}^{m} n_i$, $j \in \{1, 2, ..., n_i\}$, and $i \in \{i, 2, .., m\}$. While LDA is usually not regarded as a dimension reduction

procedure, it can be employed as one. This method is often adopted for speech and face recognition classification problems.

Despite its popularity, LDA has limitations. For example, LDA does not necessarily yield an optimal reduced subspace with dimensionality $q \ll p$ because LDA does not incorporate discriminatory information contained in the differences of the covariance matrices and parameters other than the group means. Moreover, LDA assumes that the covariance matrices for all m classes are equal, which is highly improbable in real-life problems.

Principal components analysis (PCA) is possibly the most widely used dimension reduction technique in practice. Two possible explanations for its popularity are that it is theoretically simple and that it utilizes the covariance matrices of the variables. The goal of PCA is to determine orthogonal linear combinations of the original variables, known as principal components (PCs), each with the largest possible variance. We can then reduce the data dimension by discarding the less-important principal components. The first PC is a linear combination with the largest variance. We denote the first principal component by $s_1 = \mathbf{x}'\mathbf{w}_1$, where

$$\mathbf{w}_1 = \arg \max_{\|\mathbf{w}=1\|} Var\left(\mathbf{x}'\mathbf{w}\right).$$

The second PC is the linear combination with the second largest variance that is orthogonal to the first PC, etc. The total number of PCs equals the total number of variables. In other mathematical fields, such as engineering, PCA is sometimes known as the Singular Value Decomposition (SVD), the Hotelling transform, the empirical orthogonal function (EOF) method, or the Karhunen-Loeve transform.

Despite its approval and useful properties, PCA also has shortcomings. One limitation of PCA is that it yields only a linear subspace and thus does not work well with data on nonlinear manifolds. In addition, the principal components themselves do not necessarily correspond to any meaningful physical quantities. Furthermore, the number of PCs we are required to keep is unclear, although unwritten rules

are often practiced. As a solution, one highly practiced policy is to eliminate components whose eigenvalues are smaller than a fraction of the average eigenvalue. Another rule is that we keep as many PCs as we need to explain a portion of the total variance. We can determine the number of principal components by using the eigenvalue decomposition theorem to rewrite $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}',$$

where $\mathbf{U} \in \mathbb{R}_{p\times p}$ is an orthogonal matrix containing the eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{\Lambda} = diag\,(\lambda_1, ..., \lambda_p)$ is the diagonal matrix of the ordered eigenvalues $\lambda_1 \geq ... \geq \lambda_p$. According to Fodor (2002), the total variation equals the sum of the eigenvalues of the covariance matrix such that

$$\sum_{i=1}^{p} Var\,(s_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} trace\,(\mathbf{\Sigma}),$$

where s_i denotes the i^{th} principal component and λ_i denotes the i^{th} eigenvalue. Then, the term

$$\sum_{i=1}^{k} \lambda_i / trace\,(\mathbf{\Sigma})$$

provides the cumulative proportion of explained variance for the first k principal components.

According to Martinez and Kak (2001), several differences exist between PCA and LDA. For instance, LDA focuses solely on class discrimination while PCA works directly with the data on the overall estimated covariance structure and disregards the underlying class membership. However, they also conclude that PCA outperforms LDA as a dimension reduction method when the training sample sizes are small relative to the full data dimension or when the training data does accurately represent the true distribution. In addition, Balakrishnama and Ganapathiraju (1998) claim that the shape and location of the original data sets change when transformed to a different space in PCA. On the other hand, LDA does not change

the location but simply yields more class separability while simultaneously confirming a decision region between classes.

Projection pursuit (PP) is a LDR technique different from LDA and PCA in that it incorporates third-order information, which is suitable for non-Gaussian datasets. The term "projection pursuit" was introduced by Friedman and Tukey (1974), as was the term "projection index." The projection index usually measures some countenance of non-normality because the normal distribution has the least structure. Given a projection index that defines the "interestingness" of a direction, PP searches for the directions that optimize the index. A projection is defined in the sense of "interestingness" if it contains linear or nonlinear structure. If the structure is linear, then correlations between variables are quickly detected by linear regression. Skewness, multimodality, and strong peaks are concentrated along nonlinear manifolds if the structure is nonlinear.

According to Carreira-Perpinan (1997), one primary use of PP is that it avoids the curse of dimensionality when implemented with regression or density estimation. A popular higher-order projection index is contingent upon the negative Shannon entropy. Fodor (2002) states that if we are given a random variable $\mathbf{x}$ with probability distribution f, its negative entropy is

$$Q(\mathbf{x}) \equiv \int f(\mathbf{x}) \log(f(\mathbf{x})) \, d(\mathbf{x}). \tag{1.2}$$

The normal distribution minimizes (1.2), which is in accordance with finding directions that maximize the entropy of the projected data. However, because PP works well with linear projections, it is not well-suited to deal with highly nonlinear structures. Also, computation for PP can be problematic because it handles higher than second-order information.

Factor analysis (FA) is a LDR technique that is comparable to PCA when the error terms can be assumed to have the same variance but, unlike PCA, is based on second-order data summaries. In FA, we assume that the measured variables

depend on a relatively small number of unknown common features, and we wish to discover a linear combination of original features for dimension reduction of datasets.

Like previously mentioned LDR methods, FA has advantages and disadvantages. According to Warner (2007), FA can be an expedient tool during the process of theory development and theory testing and can instruct us on how many factors are needed to account for the correlations among the variables included in the study. Moreover, FA is advantageous for the reinspection of existing measures. However, FA is often used as a desperation tactic to summarize information in a messy dataset. Also, researchers sometimes mistakenly view the results of exploratory FA as proof of the existence of latent variables, even though the set of latent factors obtained in FA is highly dependent on the selection of variables measured.

In FA, a random vector $\mathbf{x} \in \mathbb{R}_{p\times 1}$ with $E(\mathbf{x}) = \mathbf{0}$ and covariance matrix $Var(\mathbf{x}) = \mathbf{\Sigma}$ satisfies the k-factor model, where $k \leq p$, if $\mathbf{x} = \mathbf{\Lambda w} + \mathbf{v}$, where $\mathbf{\Lambda}_{p\times k}$ is a matrix of constants, and $\mathbf{w}_{k\times 1}$ and $\mathbf{v}_{p\times 1}$ are the "random common factors" and "specific factors," respectively. All factors in the model are uncorrelated, and common factors are standardized to have variance one. Additionally, $E(\mathbf{w}) = \mathbf{0}$, $Var(\mathbf{w}) = \mathbf{I}$, $E(\mathbf{v}) = \mathbf{0}$, $Cov(v_i, v_j) = 0$ for $i \neq j$, and $Cov(\mathbf{w}, \mathbf{v}) = \mathbf{0}$. Provided these assumptions hold,

$$Cov(\mathbf{v}) = \mathbf{\Psi} = diag(\psi_{11}, ..., \psi_{pp}).$$

According to Fodor (2002), if the covariance matrix is of the form $Var(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$, then the k-factor model holds. Because $x_i = \sum_{j=1}^{k} \lambda_{ij} w_j + v_i$ for $i = 1, 2, ..., p$, the variance of x_i is $\sigma_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_{ii}$, where $c_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$ is called the *communality* and represents the variance of x_i, and ψ_{ii} is called the *specific variance*, which is the contribution in the variability of x_i for its v_i. The term λ_{ij}^2 measures the importance of the dependence of x_i on the common factor w_j. If several variables x_i have high values of λ_{ij} on specified factor w_j, then the implication is that those variables are redundant measurements.

Fodor (2002) has stated that independent component analysis (ICA) is currently a popular dimension reduction method where the objective is to determine rotations that maximize standards for independence. The ICA procedure uses a higher-order model that seeks linear projections that are as nearly statistically independent as possible but not necessarily orthogonal. ICA is a generalization of the concepts behind PCA and PP, although no order exists with the ICs, contrary to PCA. While PCA looks for uncorrelated variables, where for all $i \neq j$, $1 \leq i, j \leq p$,

$$Cov\left(x_i, x_j\right) = E\left[\left(x_i - \mu_i\right)\left(x_j - \mu_j\right)\right] = E\left(x_i x_j\right) - E\left(x_i\right) E\left(x_j\right) = 0,$$

ICA looks for independent variables, where

$$f\left(x_1, ..., x_p\right) = f_1\left(x_1\right) \cdot ... \cdot f_p\left(x_p\right).$$

Independence is a much stronger property than uncorrelatedness. Uncorrelatedness involves second-order statistics, but independence depends on all the higher-order statistics. Once estimated, the independent components (ICs) can be ordered according to the norms of the columns of the mixing matrix, a similar ordering to that of PCs, or according to some non-Gaussian measure.

The noise-free ICA model for the p-dimensional random vector $\mathbf{x}$ attempts to estimate the components of the k-dimensional vector $\mathbf{s}$ and the $p \times k$ full column rank mixing matrix $\mathbf{A}_{p \times k}$, where

$$\mathbf{x} = \mathbf{A}\mathbf{s},$$

where the components of $\mathbf{s}$ are as independent as possible. Pertaining to feature selection, the columns of $\mathbf{A}$ represent the reduced feature space of the data, and the components of $\mathbf{s}$ give the reduced features.

As stated in Ravisekar (2006), ICA has several drawbacks. For instance, the transformed features we obtain while employing ICA may not be orthogonal. Also, the selection of the optimal number of dimensions of the source data vectors in ICA has yet to be addressed. Moreover, ICA might not reduce the original data

dimension. Despite its demerits, ICA has been applied to many different problems, including exploratory data analysis, blind source separation, blind deconvolution, and feature selection.

While most of the previously mentioned LDR techniques are not new, a vast amount of research involving LDR continues today. Some newer LDR methods include those by Hennig (2004), Loog and Duin (2001), Loog and Duin (2004), Khambatla and Leen (1997), Lotikar and Kothari (2000), and Tang et al. (2005).

The LDR method considered "best" depends on the specific characteristics of the dataset and the goals of the researcher. Throughout the remainder of this dissertation, we consider a preferred dimension reduction method to be the one that minimizes the Bayes probability of misclassification ($BPMC$).

1.3 Background Research for Upcoming Chapters

In this section, we focus on a LDR method first proposed by Peters, Redner, and Decell (1978) based on the concept of linear sufficient statistics. Given a dominated family, $\mathscr{D}$, of probability measures defined on the Borel subsets of a topological linear subspace X and a continuous linear transformation $T : \mathbb{R}^p \to \mathbb{R}^q$, let

$$\mathbf{M} \equiv [\boldsymbol{\mu}_2|...|\boldsymbol{\mu}_m|\boldsymbol{\Sigma}_2 - \mathbf{I}|...|\boldsymbol{\Sigma}_m - \mathbf{I}]\,. \tag{1.3}$$

Peters, Redner, and Decell (1978) have provided necessary and sufficient conditions for T to be a linear sufficient statistic for $\mathscr{D}$. This condition is $\mathbf{T}\mathbf{T}^{+}\mathbf{M} = \mathbf{M}$. Thus, they show that $rank\,(\mathbf{M}) = k$ is the smallest dimension for which there exists a linear sufficient dimension reduction matrix. However, they do not explicitly determine the $p \times k$ linear dimension reduction matrix $\mathbf{T}$.

Odell (1979) and Decell, Odell, and Coberly (1981) have constructively derived the proposed LDR method proposed by Peters, Redner, and Decell (1978) for multiple multivariate Gaussian populations with known parameters based on minimizing the $BPMC$ in the reduced space. This linear feature selection method is as

follows: Suppose we have m distinct multivariate normal populations $\Pi_1, \Pi_2, ..., \Pi_m$ with known population means $\boldsymbol{\mu}_i$ and positive-definite covariance matrices $\boldsymbol{\Sigma}_i$ with $i = 1, 2, ..., m$. Then, $q = rank\,(\mathbf{K})$ is the smallest dimension $(q \leq p)$ such that $G = G\,(\mathbf{K})$ if and only if $\mathbf{K} = \mathbf{AF}'$, where $\mathbf{M} = \mathbf{FH}$ with $rank\,(\mathbf{M}) = rank\,(\mathbf{F}) = rank\,(\mathbf{H}) = q$, $\mathbf{A}$ is an arbitrary nonsingular $q \times q$ matrix, and $\mathbf{M}$ is given in (1.3). We remark that $\mathbf{K}$ is not unique because $\mathbf{A}$ is arbitrary. Thus, if we let $\mathbf{A} = (\mathbf{F}'\mathbf{F})$, $\mathbf{K}$ becomes $\mathbf{F}^+$, the Moore-Penrose pseudoinverse of $\mathbf{F}$. Also, Odell (1979) has generalized the *LDR* problem for Gaussian populations with known parameters by showing that if the coefficients of a family of continuous stochastic processes with finite basis satisfy particular assumptions, then a q-dimensional basis may be established that contains the information in the original p-dimensional family of process for $q \ll p$.

While multivariate normal models are mathematically tractable, they do not always accurately characterize a dataset and generally are not a realistic option. For these reasons, we often depict populations to be nonnormal. Young, Odell, and Marco (1985) have shown that the *LDR* procedure proposed by Odell (1979) and Decell et al. (1981) also performs well for certain symmetric unimodal nonnormal populations without increasing the *BPMC*, and the $\mathbf{M}$ method of dimension reduction is extended to multivariate θ-generalized normal densities, a family of densities defined by Goodman and Kotz (1973).

In a realistic setting, we must estimate all population parameters, and the aforementioned *LDR* techniques, assuming known parameters, do not directly apply. Tubbs, Coberly, and Young (1982) have proposed a solution to the problem of applying the *LDR* proposed by Odell (1979) and Decell et al. (1981) when class parameters are known by using the *SVD* to determine a lower-dimensional feature space that does not significantly increase the *BPMC* for the sampling case of the Bayes classification rule. Chapters two and three of this dissertation can be viewed as an extension of the *LDR* research based on the concept of linear sufficient statistics

described in this section. Here, we apply extend the previously mentioned research to the multivariate skew-normal and the multivariate singular skew-normal distributions.

The outline of this dissertation is as follows. In Chapter two, we derive a *LDR* technique for statistical discriminant analysis with multiple multivariate skew-normal populations. We first define the multivariate skew-normal distribution, provide several real-life situations where it is applied, and detail the origin of its existence. Next, we state the distributions of a linear combination and the marginal and conditional distributional properties. Then, we assemble and prove several lemmas used in the derivation of our *LDR* method followed by a series of theorems that present our *LDR* procedure for the multivariate skew-normal populations with various restrictions on the parameter configurations. Lastly, we provide three examples illustrating our *LDR* technique for multiple multivariate skew-normal distributions.

In Chapter three, we define the multivariate singular skew-normal distribution and rigorously prove its existence. Next, we state and prove distributional properties for linear combinations, marginal random variables, and conditional random variables. Then, we state and prove several lemmas that we use in deriving our *LDR* transformation for the multiple multivariate singular skew-normal distributions with various parameter configurations. We then state and prove several theorems concerning the formulation of our *LDR* technique followed by their corresponding proofs. Finally, through two examples, we illustrate the effectiveness of our *LDR* technique for multiple multivariate singular skew-normal classes.

In Chapter four, we compare two statistical linear discrimination procedures when monotone missing data exists in the training-data sets from two different multivariate normally distributed populations with equal covariance matrices. We derive the maximum likelihood estimators (*MLE*s) for two partitioned population means and a common covariance matrix. Also, we introduce the two competing classifiers:

a linear combination discriminant function derived from Chung and Han (2000) (*C-H*) and a linear classifier based on the *MLE* of two multivariate normal training samples with identical monotone missing data in one or more features. We then use two Monte Carlo simulations with various parameter configurations to compare the utility of the *MLE* and *C-H* classifiers as the correlation increases among features for two populations with unequal means and equal covariance matrices. Also, we compare the two competing classifiers using parametric bootstrap estimated expected error rates for two real-data sets. To validate the assumption of multivariate normality, we use Mardia's test for multivariate skewness and kurtosis to establish that both populations from both data sets are multivariate normally distributed. Our Monte Carlo simulation and real-data comparison results indicate that when features are highly correlated, the *MLE* classifier can considerably outperform the *C-H* classifier, but when features are not highly correlated, the *C-H* classifier can slightly outperform the *MLE* classifier.

CHAPTER TWO

Linear Dimension Reduction for Multiple Multivariate Skew-Normal Densities

2.1 Introduction

Our goal for this chapter is to derive a linear dimension reduction method for multiple multivariate skew-normal (*MSN*) distributions. In Section 2, we define the *MSN* distribution, give its history, and present some of its useful properties. In Section 3, we provide some preliminary lemmas used in the proof of our main result. In Section 4, we give the main result, which is our linear dimension reduction technique for multiple *MSN* densities. We give three examples in Section 5 and some brief comments in Section 6.

2.2 Some Notation and the Multivariate Skew-Normal Distribution

Throughout the remainder of the chapter, we use the notation $\mathbb{R}_{m\times n}$ to represent the vector space of all $m \times n$ matrices over the real field $\mathbb{R}$. Also, we let the symbol $\mathbb{R}^S_{n\times n}$ represent the cone of all $n \times n$ symmetric matrices of real numbers. In addition, the symbols $\mathbb{R}^{\geq}_n$ and $\mathbb{R}^{>}_n$ represent the cone of all symmetric nonnegative definite and positive definite matrices, respectively, in $\mathbb{R}_{n\times n}$. Moreover, for $\mathbf{A} \in \mathbb{R}_{m\times n}$ we use $\mathcal{N}(\mathbf{A})$ to represent the null space and $\mathcal{C}(\mathbf{A})$ to denote the column space of $\mathbf{A}$.

Although the multivariate normal distribution is well-known and readily mathematically tractable, it does not always model random phenomena. The *MSN* distribution has many different applications that often appear with variables that follow properties of the normal distribution but have undergone selective reporting. When handling data, we sometimes must account for hidden truncation. For example, Arnold and Beaver (2002) provide an example where investigators collect data for

measuring and reporting archaeological pottery shards that are large enough not to fall through the sieve used on-site. Another common application for the skew-normal distribution is the modeling of insurance risk.

Azzalini (1985) introduced the multivariate skew-normal distribution as an extension of the univariate normal distribution to account for symmetry. Arnold and Beaver (2002) noticed that skew-normal distributions may be encountered in situations in which the observations comply with a normal principle but have been truncated with respect to a hidden covariable. This phenomenon was illustrated by the joint distribution of height and waist measurements of selected individuals for elite troops. In conjunction with coauthors, Azzalini extended the univariate normal distribution to include a multivariate analog of the *MSN* distribution. A formal definition of the *MSN* density function parameterized by Vernic (2006) is given below.

Definition 2.1. A random vector $\mathbf{x}$ is said to follow a MSN distribution with skewness parameter $\boldsymbol{\gamma}$, written $\mathbf{x} \sim SN_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}\right)$, if its density function is

$$p(\mathbf{x}) = \frac{1}{\Phi\left(\delta_0\right)} \varphi_p\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}}\right), \tag{2.1}$$

where $\mathbf{x} \in \mathbb{R}_{p\times 1}$, $\boldsymbol{\mu} \in \mathbb{R}_{p\times 1}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{>}_{p\times p}$, $\boldsymbol{\gamma} \in \mathbb{R}_{p\times 1}$, $\delta_0 \in \mathbb{R}$, $\phi(\mathbf{x})$ is the multivariate normal density function, and $\Phi(\mathbf{x})$ is the univariate standard normal density function.

The parameter vector $\boldsymbol{\gamma}$ regulates the skewness of (2.1), and when $\boldsymbol{\gamma} = \mathbf{0}$, density (2.1) corresponds to a multivariate normal density function. Authors such as Arnold and Beaver (2002) and Azzalini and Capitanio (1999) have often chosen $\delta_0 = 0$ to simplify the definition of the *MSN* density so that

$$p\left(\mathbf{x}\right) = 2\varphi_p\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \Phi\left(\frac{\boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}}\right).$$

Vernic (2005) has stated the moment-generating function (MGF) of density (2.1) without proof. Here, we offer a proof of the MGF of the *MSN* random vector with density (2.1).

Theorem 2.1. Let $\mathbf{v} \sim SN_p(\mathbf{0}, \mathbf{I}_p, \delta_0, \boldsymbol{\gamma})$ and $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, where $\mathbf{v} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu})$. Hence,

$$M_{\mathbf{v}}(\mathbf{t}) = \frac{1}{\Phi(\delta_0)} \exp\left\{\frac{\mathbf{t}'\mathbf{t}}{2}\right\} \Phi\left(\frac{\lambda_0 + \boldsymbol{\lambda}_1'\mathbf{t}}{\sqrt{1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}\right),$$

where

$$\delta_0 \equiv \frac{\lambda_0}{\sqrt{1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}, \boldsymbol{\lambda}_1 \equiv \frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\gamma}}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}, \text{ and } 1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1 = 1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}$$

with

$$\lambda_0 \equiv \frac{\delta_0}{\sqrt{1 + \boldsymbol{\delta}_1'\boldsymbol{\delta}_1}} \text{ and } \boldsymbol{\delta}_1 \equiv \frac{\boldsymbol{\lambda}_1}{\sqrt{1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}.$$

Also, note that

$$\begin{aligned} \frac{\boldsymbol{\lambda}_1'\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{\sqrt{1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}} &= \frac{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}{1} \frac{\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}} \\ &= \boldsymbol{\gamma}'\mathbf{t}. \end{aligned} \tag{2.2}$$

Then,

$$\begin{aligned} M_{\mathbf{x}}(\mathbf{t}) &= M_{\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{v}}(\mathbf{t}) \\ &= \exp\{\mathbf{t}'\boldsymbol{\mu}\} M_{\mathbf{V}}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{t}\right) \\ &= \exp\{\mathbf{t}'\boldsymbol{\mu}\} \frac{1}{\Phi(\delta_0)} \exp\left\{\frac{\mathbf{t}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{2}\right\} \Phi\left(\frac{\lambda_0 + \boldsymbol{\lambda}_1'\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{\sqrt{1 + \boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}\right) \\ &= \frac{1}{\Phi(\delta_0)} \exp\left\{\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\right\} \Phi\left(\delta_0 + \frac{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}} \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}\right) \\ &= \exp\left\{\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\right\} \frac{\Phi(\delta_0 + \boldsymbol{\gamma}'\mathbf{t})}{\Phi(\delta_0)}. \end{aligned}$$

Propositions pertaining to a linear combination MSN of random variables, the marginal MSN distribution for a MSN random vector, and the conditional distribution of components of a MSN random vector originally provided by Vernic (2005) are given below. The following proposition expresses the distribution of $\mathbf{Cx} + \mathbf{b}$, where the vector $\mathbf{x}$ is MSN, $\mathbf{C} \in \mathbb{R}_{m \times p}$ is full row rank, and $\mathbf{b} \in \mathbb{R}_{m \times 1}$.

Proposition 2.1. Let $\mathbf{b} \in \mathbb{R}_{m\times 1}$ and $\mathbf{C} \in \mathbb{R}_{m\times p}$ with rank m, where $m \leq p$. If $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, then $\mathbf{b} + \mathbf{C}\mathbf{x} \sim SN_p(\mathbf{b} + \mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}', \delta_0, \mathbf{C}\boldsymbol{\gamma})$.

The following proposition describes the properties of subvectors of a *MSN* vector.

Proposition 2.2. Let $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$. If we partition $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ into two subvectors of dimensions m and $p-m$, respectively, and correspondingly partition

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \text{ and } \boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix}, \tag{2.3}$$

then

(i) $\mathbf{x}_1 \sim SN_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \delta_0, \boldsymbol{\gamma}_1)$;

(ii) $\mathbf{x}_2 \sim SN_{p-m}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}, \delta_0, \boldsymbol{\gamma}_2)$.

In particular, the univariate marginal distributions of $\mathbf{x}$ are given by

$$x_j \sim SN_1\left(\mu_j, \sigma_j^2, \delta_0, \gamma_j\right)$$

with $\sigma_j^2 = \sigma_{jj}$ for $j = 1, ..., p$.

The above proposition regarding the marginal distribution of a *MSN* random vector is now used for deriving the consequent proposition regarding the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$. Vernic (2005) references Arnold and Beaver (2002) to express the conditional density function of a partitioned *MSN* density function as stated in Proposition 2.2. We present a proof to validate our assertion that the conditional density function of a *MSN* density function does not follow a skew-normal distribution. We use a lemma from Vernic (2005) to derive our result.

Lemma 2.1. Let $\boldsymbol{\Sigma} \in \mathbb{R}_{p\times p}$, $\boldsymbol{\mu} \in \mathbb{R}_{p\times 1}$, and $\boldsymbol{\gamma} \in \mathbb{R}_{p\times 1}$ be defined as in (2.3). Then,

$$\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) + \left(\boldsymbol{\gamma}_1' - \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)\boldsymbol{\Sigma}_{1|2}^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right).$$

Proof: First, Vernic (2005) has noted that

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix},$$

where $\mathbf{T}_{11} = \boldsymbol{\Sigma}_{1|2}^{-1} = \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1}$, $\mathbf{T}_{12} = -\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{2|1}^{-1}$, $\mathbf{T}_{21} = -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{1|2}^{-1}$, and $\mathbf{T}_{22} = \boldsymbol{\Sigma}_{2|1}^{-1} = \left(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11\boldsymbol{\Sigma}_{12}}^{-1}\right)^{-1}$ and

$$\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} = -\mathbf{T}_{11}^{-1}\mathbf{T}_{12}.$$

Then,

$$\begin{aligned}
\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= \left(\boldsymbol{\gamma}_1'\mathbf{T}_{11} + \boldsymbol{\gamma}_2'\mathbf{T}_{21}\right)(\mathbf{x}_1 - \boldsymbol{\mu}_1) + \left(\boldsymbol{\gamma}_1'\mathbf{T}_{12} + \boldsymbol{\gamma}_2'\mathbf{T}_{22}\right)(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \left(\boldsymbol{\gamma}_1'\mathbf{T}_{11} + \boldsymbol{\gamma}_2'\mathbf{T}_{21}\right)\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right) + \left[\left(\boldsymbol{\gamma}_1'\mathbf{T}_{11} + \boldsymbol{\gamma}_2'\mathbf{T}_{21}\right)\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\right. \\
&\quad \left. + \left(\boldsymbol{\gamma}_1'\mathbf{T}_{12} + \boldsymbol{\gamma}_2'\mathbf{T}_{22}\right)\right](\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \left(\boldsymbol{\gamma}_1'\mathbf{T}_{11} + \boldsymbol{\gamma}_2'\mathbf{T}_{21}\right)\mathbf{T}_{11}^{-1}\mathbf{T}_{11}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right) + \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\
&= \left(\boldsymbol{\gamma}_1' - \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)\boldsymbol{\Sigma}_{1|2}^{-1}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right) + \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2).
\end{aligned}$$

Proposition 2.3. Let $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, and let the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\gamma}$ be partitioned as in (2.3). Then, the conditional density function of $\mathbf{x}_1|\mathbf{x}_2$ is

$$\frac{1}{\Phi(\beta_0)}\varphi_m\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right)\Phi\left(l_0 + \mathbf{l}_1'\boldsymbol{\Sigma}_{1|2}^{-\frac{1}{2}}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right)\right),$$

where

$$\boldsymbol{\mu}_{1|2} \equiv \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \tag{2.4}$$

$$\boldsymbol{\Sigma}_{1|2} \equiv \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}, \tag{2.5}$$

$$l_0 \equiv \frac{\delta_0 + \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}, \tag{2.6}$$

$$\beta_0 \equiv \frac{\delta_0 + \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)}{\sqrt{1 - \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\gamma}_2}}, \tag{2.7}$$

and

$$\mathbf{l}_1 \equiv \frac{\left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1/2}\left(\boldsymbol{\gamma}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\gamma}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}. \tag{2.8}$$

Proof: From Definition 2.1, the joint probability density function of $\mathbf{x}$ is

$$f(\mathbf{x}) = \frac{1}{\Phi(\delta_0)} \varphi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}} \right).$$

Also, Proposition 2.2 states that the marginal density function of $\mathbf{x}_2$ is

$$f(\mathbf{x}_2) = \frac{1}{\Phi(\delta_0)} \varphi_{p-m}(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)}{\sqrt{1 - \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\gamma}_2}} \right).$$

Then by Lemma 2.1, the conditional density function of $\mathbf{x}_1 | \mathbf{x}_2$ is

$$\begin{aligned}
f(\mathbf{x}_1 | \mathbf{x}_2) &= \frac{\frac{1}{\Phi(\delta_0)} \varphi_p(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}} \right)}{\frac{1}{\Phi(\delta_0)} \varphi_{p-m}(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)}{\sqrt{1 - \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\gamma}_2}} \right)} \\
&= \frac{1}{\Phi(\beta_0)} \varphi_m\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}} \right) \\
&= \frac{1}{\Phi(\beta_0)} \varphi_m\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right) \times \\
&\quad \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_{22}' \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}} + \frac{\left(\boldsymbol{\gamma}_1 - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\gamma}_2\right)'}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}}} \boldsymbol{\Sigma}_{1|2}^{-1} \left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right) \right) \\
&= \frac{1}{\Phi(\beta_0)} \varphi_m\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right) \Phi\left(l_0 + \mathbf{l}_1' \boldsymbol{\Sigma}_{1|2}^{-\frac{1}{2}} \left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right) \right), \qquad (2.9)
\end{aligned}$$

where $\boldsymbol{\mu}_{1|2}$, $\boldsymbol{\Sigma}_{1|2}$, l_0, β_0, and $\mathbf{l}_1$ are defined in (2.4) - (2.8), respectively. Clearly, (2.9) is not in the form of the skew-normal density function in Definition 2.1, which implies that the conditional density function of a *MSN* distribution does not have a *MSN* density function.

2.3 Bayes Statistical Classification

The Bayes statistical classifier discriminates based on knowledge of the probability density functions $p(\mathbf{x}|\Pi_i)$, $i = 1, 2 \ldots, m$, of each class. Bayes discrimination is optimal in the sense that it maximizes the class *a posteriori* probability provided all class distributions are known. More precisely, suppose we have m classes $\Pi_1, \Pi_2, ..., \Pi_m$ with *a priori* probabilities $\alpha_1, \alpha_2, \ldots, \alpha_m$ that are assumed known. Let $\lambda(\Pi_i|\Pi_j)$ be a measure of loss when $\mathbf{x}$ is assigned to class Π_i and belongs to

class Π_j, $i \neq j$. The goal of statistical decision theory is to obtain a decision rule that assigns an unlabeled observation $\mathbf{x}$ to Π_k if $p(\Pi_k|\mathbf{x})$ is the maximum overall *a posteriori* density $p(\Pi_i|\mathbf{x}),\ i = 1, 2, \ldots, m$.

More precisely, if one assumes the loss function

$$\lambda\left(\Pi_i|\Pi_j\right) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases},$$

the Bayes classifier assigns $\mathbf{x}$ to class Π_k if

$$p(\Pi_k|\mathbf{x}) > p(\Pi_j|\mathbf{x}),\ \ j = 1, 2, \ldots, m,\ \ j \neq k.$$

This decision rule partitions the measurement or feature space into m disjoint regions $\Pi_1, \Pi_2, \ldots, \Pi_m$ such that $\mathbf{x}$ is assigned to class Π_k if $\mathbf{x} \in \Pi_k$. Using Bayes' rule, one can express the *a posteriori* probabilities of class membership $p(\Pi_k|\mathbf{x})$ as

$$p(\Pi_k|\mathbf{x}) = \frac{\alpha_k p(\mathbf{x}|\Pi_k)}{p(\mathbf{x})}.$$

Then, one can re-express the Bayes classification as:

Assign $\mathbf{x}$ to Π_k if

$$\alpha_k p(\mathbf{x}|\Pi_k) > \alpha_j\, p(\mathbf{x}|\Pi_j)\ j = 1, \ldots, m,\ \ j \neq k.$$

This decision rule is known as Bayes' classification rule for *minimum error*.

We now consider the assumption that the class density $p(\mathbf{x}|\Pi_i)$ has a *MSN* density function, i.e., $\mathbf{x} \sim SN_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i\right),\ i \in \{1, 2, ..., m\}$, where $\boldsymbol{\mu}_i,\ \boldsymbol{\gamma}_i \in \mathbb{R}_{p\times 1}$ and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{>}_{p\times p}$. Thus, for class Π_i, $\mathbf{x}$ has density

$$p(\mathbf{x}|\Pi_i) = \frac{1}{\Phi\left(\delta_0\right)} \varphi_p\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_i' \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i}{\sqrt{1 - \boldsymbol{\gamma}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\gamma}_i}}\right).$$

Because (2.1) is positive and the logarithm function is monotonic, the Bayes classifier for classifying $\mathbf{x}$ into one of two *MSN* densities can be expressed as:

Assign $\mathbf{x}$ to Π_1 if

$$-\ln\left(\Phi\left(\delta_0\right)\right) + \ln\left(\varphi_p\left(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)\right) + \ln\left(\Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_1' \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \boldsymbol{\mu}_1)}{\sqrt{1 - \boldsymbol{\gamma}_1' \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\gamma}_1}}\right)\right) >$$

$$-\ln\left(\Phi\left(\delta_0\right)\right)+\ln\left(\varphi_p\left(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2\right)\right)+\ln\left(\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}_2'\,\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}{\sqrt{1-\boldsymbol{\gamma}_2'\,\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\gamma}_2}}\right)\right)$$

and to Π_2, otherwise.

For the case of $m>2$ populations or classes, the generalized distance function is

$$d_i(\mathbf{x})\equiv-\ln\left(\Phi\left(\delta_0\right)\right)+\ln\left(\varphi_n\left(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\right)\right)+\ln\left(\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}_i'\,\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{\sqrt{1-\boldsymbol{\gamma}_i'\,\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\gamma}_i}}\right)\right),$$

where $i=1,2,...,m$. The Bayes decision rule is to classify an unlabeled observation vector $\mathbf{x}$ into the class Π_k if $d_k(\mathbf{x})=\min\{d_i(\mathbf{x}),\ i=1,2,\ldots,m\}$.

2.4 Preliminary Results

The proof of our new linear dimension reduction theorem requires the following notation and lemmas. Consider $\mathbf{M}\in\mathbb{R}_{p\times(m-1)(p+1)}$, where

$$\mathbf{M}\equiv\left[\mathbf{E}_2^{-1}\left(\mathbf{d}_2-\mathbf{d}_1\right)|...|\mathbf{E}_m^{-1}\left(\mathbf{d}_m-\mathbf{d}_1\right)|\mathbf{E}_2-\mathbf{E}_1|...|\mathbf{E}_m-\mathbf{E}_1|\mathbf{a}_2-\mathbf{a}_1|...|\mathbf{a}_m-\mathbf{a}_1\right],$$

with $\mathbf{a}_i,\mathbf{d}_i\in\mathbb{R}_{p\times1}$, $\mathbf{E}_i\in\mathbb{R}_{p\times p}^{S}$, where $rank\left(\mathbf{E}_i\right)=p$ for $i=1,2,...,m$, and $\mathbf{E}_1\neq\mathbf{E}_k$ for some k, where $2\leq k\leq m$. Also, let $rank(\mathbf{M})=q<p$, and let $\mathbf{M}=\mathbf{FG}$, where $\mathbf{F}\in\mathbb{R}_{p\times q}$ and $\mathbf{G}\in\mathbb{R}_{q\times(m-1)(p+1)}$ with $rank\left(\mathbf{F}\right)=rank\left(\mathbf{G}\right)=q$. Then, the Moore-Penrose pseudoinverse of $\mathbf{M}$ is $\mathbf{M}^{+}=\mathbf{G}^{+}\mathbf{F}^{+}$, $\mathbf{MM}^{+}=\mathbf{FGG}^{+}\mathbf{F}^{+}=\mathbf{FF}^{+}$, and $\mathbf{MM}^{+}\mathbf{M}=\mathbf{FF}^{+}\mathbf{M}=\mathbf{M}$. This property implies that for $i=1,2,...,m$,

(1) $\mathbf{FF}^{+}\left(\mathbf{a}_i-\mathbf{a}_1\right)=\mathbf{a}_i-\mathbf{a}_1$,

(2) $\mathbf{FF}^{+}\left(\mathbf{E}_i-\mathbf{E}_1\right)=\mathbf{E}_i-\mathbf{E}_1$,

(3) $\mathbf{FF}^{+}\left[\mathbf{E}_i^{-1}\left(\mathbf{d}_i-\mathbf{d}_1\right)\right]=\mathbf{E}_i^{-1}\left(\mathbf{d}_i-\mathbf{d}_1\right)$.

We now provide nine lemmas that we use in the proof of our main *LDR* results. The reader can find the proofs of Lemmas 2.1-2.3 in Onsupreth and Young (2008).

Lemma 2.2. Let $\mathbf{d}_i\in\mathbb{R}_{p\times1}$, $\mathbf{F}\in\mathbb{R}_{p\times q}$, and $\mathbf{C}=\mathbf{R}[\mathbf{I}-\mathbf{FF}^{+}]$, where $\mathbf{R}\in\mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{C})=p-q$ and $\mathbf{E}_i\in\mathbb{R}_{p\times p}^{S}$ for $i=1,2,...,m$, such that properties

(1) - (3) hold. Then,

(a) $\mathbf{F}\mathbf{F}^{+}(\mathbf{E}_i - \mathbf{E}_1) = (\mathbf{E}_i - \mathbf{E}_1)\mathbf{F}\mathbf{F}^{+}$,

(b) $\mathbf{F}\mathbf{F}^{+}\mathbf{E}_i = \mathbf{E}_i\mathbf{F}\mathbf{F}^{+}$,

(c) $(\mathbf{I} - \mathbf{F}\mathbf{F}^{+})\mathbf{E}_i = \mathbf{E}_1(\mathbf{I} - \mathbf{F}\mathbf{F}^{+})$,

(d) $\mathbf{C}\mathbf{E}_i\mathbf{C}' = \mathbf{C}\mathbf{E}_1\mathbf{C}'$.

The following lemma establishes the inverse of the quantity $\mathbf{F}^{+}\mathbf{\Sigma}_i\mathbf{F}^{+\prime}$.

Lemma 2.3. Let $\mathbf{F} \in \mathbb{R}_{p\times q}$ and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\,(\mathbf{E}_i) = p$ for $i = 1, ..., m$, such that properties (1) - (3) hold. Then, $(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime})^{-1} = \mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F}$.

The subsequent lemma gives the inverse of a full-rank partitioned matrix, and one can find a proof in Lewis and Odell (1971).

Lemma 2.4. Let $\mathbf{A} \in \mathbb{R}_{p\times p}$, where $rank\,(\mathbf{A}) = p$, and let $\mathbf{A}$ be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix},$$

where $\mathbf{A}_{11} \in \mathbb{R}_{q\times q}$, $\mathbf{A}_{22} \in \mathbb{R}_{(p-q)\times(p-q)}$, $\mathbf{A}_{21} \in \mathbb{R}_{(p-q)\times q}$, and $\mathbf{A}_{12} \in \mathbb{R}_{q\times(p-q)}$. Then,

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix},$$

where $\mathbf{B}_{11} = \left[\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right]^{-1}$, $\mathbf{B}_{22} = \left[\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right]^{-1}$,
$\mathbf{B}_{12} = -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{B}_{22}$, and $\mathbf{B}_{21} = -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{B}_{11}$.

The following lemma gives results that are used in the proof of our main dimension reduction theorem.

Lemma 2.5. Let $\mathbf{a}_i, \mathbf{d}_i \in \mathbb{R}_{p\times 1}$, $\mathbf{F} \in \mathbb{R}_{p\times q}$, and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\,(\mathbf{E}_i) = p$ for $i = 1, ..., m$, such that properties (1) - (3) hold, and let $\mathbf{C} = \mathbf{R}[\mathbf{I} - \mathbf{F}\mathbf{F}^{+}]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{C}) = p - q$. Then, for $i = 2, \ldots, m$,

(a) $\mathbf{CF} = \mathbf{0}$,

(b) $\mathbf{Cd}_i = \mathbf{Cd}_1$,

(c) $\mathbf{Ca}_i = \mathbf{Ca}_1$,

(d) $\mathbf{Cd}_i + \mathbf{CE}_i\mathbf{F}^{+\prime}(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime})^{-1}(\mathbf{y} - \mathbf{F}^{+}\ \mathbf{d}_i) = \mathbf{Cd}_1$, where $\mathbf{y} \in \mathbb{R}_{p\times 1}$,

(e) $\mathbf{CE}_i\mathbf{C}' - \mathbf{CE}_i\mathbf{F}^{+\prime}(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime})^{-1}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' = \mathbf{CE}_1\mathbf{C}'$,

(f) $\mathbf{a}_i'\mathbf{C}' \left[\mathbf{CE}_i\mathbf{C}' - \mathbf{CE}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right]^{-1}(\mathbf{Cx} - \mathbf{Cd}_1)$

$= (\mathbf{Ca}_1)'\left[\mathbf{CE}_1\mathbf{C}'\right]^{-1}(\mathbf{Cx} - \mathbf{Cd}_1)$,

(g) $(\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}')\left[\mathbf{CE}_i\mathbf{C}' - \mathbf{CE}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right]^{-1}\mathbf{Ca}_i = \mathbf{0}$,

(h) $\mathbf{Ca}_i - \mathbf{CE}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\mathbf{F}^{+}\mathbf{a}_i = \mathbf{Ca}_1$.

Proof of (a). The proof of (a) is trivial and is thus omitted.

Proof of (b). Let $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$. From property (3), we have that

$$\begin{aligned}\mathbf{FF}^{+}\mathbf{E}_i^{-1}(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{E}_i^{-1}(\mathbf{d}_i - \mathbf{d}_1) &\Rightarrow \mathbf{E}_i^{-1}\mathbf{FF}^{+}(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{E}_i^{-1}(\mathbf{d}_i - \mathbf{d}_1)\\ &\Rightarrow \left(\mathbf{I} - \mathbf{FF}^{+}\right)(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{0}\\ &\Rightarrow \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^{+}\right)(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{0}\\ &\Rightarrow \mathbf{Cd}_i = \mathbf{Cd}_1.\end{aligned}$$

Proof of (c). From property (1), we have that

$$\begin{aligned}\mathbf{FF}^{+}(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{a}_i - \mathbf{a}_1 &\\ \Rightarrow \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^{+}\right)(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{0} &\\ \Rightarrow \mathbf{Ca}_i = \mathbf{Ca}_1.&\end{aligned}$$

Proof of (d). First, note that

$$\begin{aligned}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' &= \mathbf{F}^{+}\mathbf{E}_i\left(\mathbf{I}-\mathbf{F}\mathbf{F}^{+}\right)\mathbf{R}' \\ &= \left(\mathbf{F}^{+}\mathbf{E}_i-\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}\mathbf{F}^{+}\right)\mathbf{R}' \\ &= \mathbf{0}.\end{aligned}$$

Because $\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime} = \left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right)'$, the term $\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\left(\mathbf{y}-\mathbf{F}^{+}\mathbf{d}_i\right) = 0$. Then, by Lemma 2.5.b, we conclude that

$$\mathbf{C}\mathbf{d}_i + \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\left(\mathbf{y}-\mathbf{F}^{+}\mathbf{d}_i\right) = \mathbf{C}\mathbf{d}_1.$$

Proof of (e). Part (e) follows from Lemma 2.3 because

$$\begin{aligned}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime})^{-1}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' &= \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' \\ &= \mathbf{C}\mathbf{E}_i\mathbf{E}_i^{-1}\mathbf{F}\mathbf{F}^{+}\mathbf{F}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' \\ &= \mathbf{C}\mathbf{F}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' \\ &= \mathbf{0}.\end{aligned}$$

Proof of (f). Using Lemmas 2.5.c and 2.5.e, we have

$$\begin{aligned}&\left(\mathbf{C}\mathbf{a}_i\right)'\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'-\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right]^{-1}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\mathbf{d}_1\right) \\ &\quad= \left(\mathbf{C}\mathbf{a}_1\right)'\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'-\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\mathbf{E}_i^{-1}\mathbf{E}_i\mathbf{F}\mathbf{F}^{+}\mathbf{C}'\right]^{-1}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\mathbf{d}_1\right) \\ &\quad= \left(\mathbf{C}\mathbf{a}_1\right)'\left[\mathbf{C}\mathbf{E}_1\mathbf{C}'\right]^{-1}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\mathbf{d}_1\right).\end{aligned}$$

Proof of (g). Using Lemma 2.5.e, we see that

$$\begin{aligned}&\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right)\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'-\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right]^{-1}\mathbf{C}\mathbf{a}_i \\ &\quad= \left[\mathbf{F}^{+}\mathbf{E}_i\left(\mathbf{I}-\mathbf{F}\mathbf{F}^{+}\right)\mathbf{R}'\right]\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'\right]^{-1}\mathbf{C}\mathbf{a}_i \\ &\quad= \left[\mathbf{F}^{+}\left(\mathbf{I}-\mathbf{F}\mathbf{F}^{+}\right)\mathbf{E}_i\mathbf{R}'\right]\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'\right]^{-1}\mathbf{C}\mathbf{a}_i \\ &\quad= \mathbf{0}.\end{aligned}$$

Proof of (h). Using Lemma 2.2.b, Lemma 2.3, Lemma 2.5.b, and the fact that $\mathbf{FF}^{+}$ is orthogonal to $\mathbf{C}$, we have

$$\begin{aligned}\mathbf{Ca}_i - \mathbf{CE}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\mathbf{F}^{+}\mathbf{a}_i &= \mathbf{Ca}_i - \mathbf{CE}_i\mathbf{F}^{+\prime}\left(\mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F}\right)\mathbf{F}^{+}\mathbf{a}_i \\ &= \mathbf{Ca}_1 - \mathbf{CE}_i\left(\mathbf{FF}^{+}\right)'\mathbf{E}_i^{-1}\mathbf{FF}^{+}\mathbf{a}_i \\ &= \mathbf{Ca}_1 - \mathbf{C}\left(\mathbf{FF}^{+}\right)'\mathbf{E}_i\mathbf{E}_i^{-1}\mathbf{FF}^{+}\mathbf{a}_i \\ &= \mathbf{Ca}_1.\end{aligned}$$

Lemma 2.6. Let $\mathbf{F} \in \mathbb{R}_{p\times q}$ and $\mathbf{C} = \mathbf{R}[\mathbf{I} - \mathbf{FF}^{+}]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{C}) = p - q$, and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ for $i = 1, ..., m$, such that properties (1) - (3) hold, and let

$$\mathbf{A} = \begin{bmatrix} \mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime} & \mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' \\ \mathbf{CE}_i\mathbf{F}^{+\prime} & \mathbf{CE}_i\mathbf{C}' \end{bmatrix}.$$

Then,

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{CE}_1\mathbf{C}'\right)^{-1} \end{bmatrix}.$$

Proof: Let $\mathbf{A}$ be defined in the statement of the lemma. Then, by Lemma 2.4,

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix},$$

where

$$\begin{aligned}\mathbf{B}_{22} &= \left[\mathbf{CE}_i\mathbf{C}' - \left(\mathbf{CE}_i\mathbf{C}'\right)\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right)\right]^{-1} \\ &= \left[\mathbf{CE}_1\mathbf{C}'\right]^{-1}\end{aligned}$$

by Lemmas 2.5.e and 2.2.c. Next, by Lemmas 2.2.b and 2.3, and because $\mathbf{FF}^{+} \in \mathcal{C}^{\perp}(\mathbf{C})$, we have

$$\begin{aligned}\mathbf{B}_{12} &= \left[-\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}^{\prime}\right)\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1}\right] \\ &= -\mathbf{F}^{\prime}\mathbf{E}_i^{-1}\mathbf{F}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}^{\prime}\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1} \\ &= -\mathbf{F}^{\prime}\mathbf{E}_i^{-1}\mathbf{E}_i\mathbf{F}\mathbf{F}^{+}\mathbf{C}^{\prime}\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1} \\ &= \mathbf{0}.\end{aligned}$$

Next, by Lemma 2.2.b and Lemma 2.3, we see that

$$\begin{aligned}\mathbf{B}_{11} &= \left[\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime} - \mathbf{F}^{+}\mathbf{E}_i\mathbf{C}^{\prime}\left(\mathbf{C}\mathbf{E}_i\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\right]^{-1} \\ &= \left[\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime} - \mathbf{F}^{+}\mathbf{E}_i\left(\mathbf{I}-\mathbf{F}\mathbf{F}^{+}\right)\mathbf{R}^{\prime}\left(\mathbf{C}\mathbf{E}_i\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\right]^{-1} \\ &= \left[\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime} - \left[\mathbf{F}^{+}\mathbf{E}_i - \mathbf{F}^{+}\mathbf{E}_i\mathbf{F}\mathbf{F}^{+}\right]\mathbf{R}^{\prime}\left(\mathbf{C}\mathbf{E}_i\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\right]^{-1} \\ &= \left[\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right]^{-1} \\ &= \mathbf{F}^{\prime}\mathbf{E}_i^{-1}\mathbf{F}.\end{aligned}$$

Finally, by Lemma 2.2.b and because $\mathbf{C} \notin \mathcal{C}\left(\mathbf{F}\mathbf{F}^{+}\right)$, we have that

$$\begin{aligned}\mathbf{B}_{21} &= \left[-\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{-1}\right] \\ &= -\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\mathbf{E}_i^{-1}\mathbf{F}\right) \\ &= -\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{E}_i\mathbf{E}_i^{-1}\left(\mathbf{F}\mathbf{F}^{+}\right)^{\prime}\mathbf{F} \\ &= -\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{F} \\ &= \mathbf{0}.\end{aligned}$$

Hence, Lemma 2.6 holds.

Lemma 2.7. Let $\mathbf{a}_i \in \mathbb{R}_{p\times 1}$, $\mathbf{F} \in \mathbb{R}_{p\times q}$, and $\mathbf{G} \in \mathbb{R}_{q\times(m-1)(p+1)}$, and let $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}^{\prime} \end{bmatrix}^{\prime}$ be a full-rank matrix, where $\mathbf{C} = \mathbf{R}[\mathbf{I} - \mathbf{F}\mathbf{F}^{+}]$ and $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{C}) = p - q$. Also, let $rank(\mathbf{M}) = q < p$, and let $\mathbf{F}$ and $\mathbf{G}$ be matrix components of a full-rank decomposition of $\mathbf{M}$ so that $\mathbf{M} = \mathbf{F}\mathbf{G}$ with $rank(\mathbf{F}) = rank(\mathbf{G}) = q$. Then, for $i = 2, \ldots, m$, we have

$$\left(\mathbf{H}\mathbf{a}_i\right)^{\prime}\mathbf{H}\mathbf{E}_i^{-1}\mathbf{H}^{\prime}\left(\mathbf{H}\mathbf{a}_i\right) = \mathbf{a}_i^{\prime}\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\mathbf{E}_i^{-1}\mathbf{F}\right)\mathbf{F}^{+}\mathbf{a}_i + \mathbf{a}_1^{\prime}\mathbf{C}^{\prime}\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{-1}\mathbf{C}\mathbf{a}_1.$$

Proof: We have that

$$
\begin{aligned}
(\mathbf{Ha}_i)' \mathbf{H}\mathbf{E}_i^{-1}\mathbf{H}' (\mathbf{Ha}_i) &= \mathbf{a}_i'\mathbf{H}' \begin{bmatrix} \mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}\mathbf{E}_1\mathbf{C}')^{-1} \end{bmatrix} \mathbf{Ha}_i \\
&= \begin{bmatrix} \mathbf{a}_i'\mathbf{F}^{+\prime} & \mathbf{a}_i'\mathbf{C}' \end{bmatrix} \begin{bmatrix} \mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & (\mathbf{C}\mathbf{E}_1\mathbf{C}')^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{F}^+\mathbf{a}_i \\ \mathbf{C}\mathbf{a}_i \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{a}_i'\mathbf{F}^{+\prime}\mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F} & \mathbf{a}_i'\mathbf{C}' (\mathbf{C}\mathbf{E}_1\mathbf{C}')^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{F}^+\mathbf{a}_i \\ \mathbf{C}\mathbf{a}_i \end{bmatrix} \\
&= \mathbf{a}_i'\mathbf{F}^{+\prime}\mathbf{F}'\mathbf{E}_i^{-1}\mathbf{F}\mathbf{F}^+\mathbf{a}_i + \mathbf{a}_1'\mathbf{C}' (\mathbf{C}\mathbf{E}_1\mathbf{C}')^{-1} \mathbf{C}\mathbf{a}_1.
\end{aligned}
$$

Lemma 2.8. Let $\mathbf{x} \sim SN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$ for $i \in \{1, 2, ..., m\}$, let $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$, where $\mathbf{F}$ and $\mathbf{C}$ are defined in Lemma 2.2, and let $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$. Then,

$$
\begin{aligned}
g\left(\mathbf{F}^+\mathbf{x} | \mathbf{C}\mathbf{x}; \Pi_i\right) = \frac{1}{\Phi(\delta_0)} \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \times \\
\Phi\left(l_{0i} + \mathbf{l}_{1i}'\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right),
\end{aligned} \tag{2.10}
$$

where

$$
l_{0i} \equiv \frac{\delta_0}{\sqrt{1-k_i}}, \tag{2.11}
$$

and

$$
\mathbf{l}_{1i} \equiv \frac{\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-1/2}\left(\mathbf{F}^+\boldsymbol{\gamma}_i\right)}{\sqrt{1-k_i}}, \tag{2.12}
$$

with

$$
k_i = \boldsymbol{\gamma}_i'\mathbf{F}^{+\prime}\mathbf{F}'\boldsymbol{\Sigma}_i^{-1}\mathbf{F}\mathbf{F}^+\boldsymbol{\gamma}_i.
$$

Proof: Using Proposition 2.2, Proposition 2.3, and the fact that $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$, we have

$$
\begin{aligned}
l_{0i} &= \frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}_\mathbf{1}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1-k_i}} \\
&= \frac{\delta_0}{\sqrt{1-k_i}}.
\end{aligned}
$$

Also, because $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$,

$$\frac{\delta_0 + (\mathbf{C}\boldsymbol{\gamma}_1)' (\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}')^{-1} (\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1)}{\sqrt{1 - (\mathbf{C}\boldsymbol{\gamma}_1)' (\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}')^{-1} (\mathbf{C}\boldsymbol{\gamma}_1)}} = \delta_0.$$

In addition, from Proposition 2.3 the MSN location parameter for (2.10) is

$$\begin{aligned}
&\mathbf{F}^+\boldsymbol{\mu}_i + \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{C}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} (\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1) \\
&\qquad = \mathbf{F}^+\boldsymbol{\mu}_i + \mathbf{F}^+\boldsymbol{\Sigma}_i \left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right) \mathbf{R}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} (\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1) \\
&\qquad = \mathbf{F}^+\boldsymbol{\mu}_i + \left(\mathbf{F}^+\boldsymbol{\Sigma}_i - \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}\mathbf{F}^+\right) \mathbf{R}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \\
&\qquad = \mathbf{F}^+\boldsymbol{\mu}_i,
\end{aligned}$$

and the corresponding dispersion parameter is

$$\begin{aligned}
&\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} - \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{C}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} \\
&\qquad = \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} - \mathbf{F}^+\boldsymbol{\Sigma}_i \left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right) \mathbf{R}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} \\
&\qquad = \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} - \left(\mathbf{F}^+\boldsymbol{\Sigma}_i - \mathbf{F}^+\boldsymbol{\Sigma}_i\right) \mathbf{R}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} \\
&\qquad = \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}.
\end{aligned}$$

Finally, Proposition 2.3 gives

$$\mathbf{l}_{1i} \equiv \frac{\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} - \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{C}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^+\boldsymbol{\gamma}_i - \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{C}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \mathbf{C}\boldsymbol{\gamma}_1\right)}{\sqrt{1 - k_i}},$$

which simplifies to

$$\mathbf{l}_{1i} = \frac{\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^+\boldsymbol{\gamma}_i\right)}{\sqrt{1 - k_i}}.$$

In addition, note that

$$\begin{aligned}
k_i &= \boldsymbol{\gamma}_1'\mathbf{C}' \left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1} \mathbf{C}\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_i'\mathbf{F}^{+\prime}\mathbf{F}'\boldsymbol{\Sigma}_i^{-1}\mathbf{F}\mathbf{F}^+\boldsymbol{\gamma}_i \\
&= \boldsymbol{\gamma}_i'\mathbf{F}^{+\prime}\mathbf{F}'\boldsymbol{\Sigma}_i^{-1}\mathbf{F}\mathbf{F}^+\boldsymbol{\gamma}_i
\end{aligned}$$

because $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$.

Lemma 2.9. Define the random vector $\mathbf{x} \sim SN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$ for class Π_i, where $i \in \{1, 2, ..., m\}$, and let $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$ be a full-rank matrix, where $\mathbf{F}^+$ and $\mathbf{C}$

are defined in Lemma 2.2. Then, $\mathbf{Cx} \sim SN_{p-q}\left(\mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}', \delta_0, \mathbf{C}\boldsymbol{\gamma}_1\right)$, where

$$f\left(\mathbf{Cx}|\Pi_i\right) = \frac{1}{\Phi\left(\delta_0\right)}\varphi_{p-q}\left(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)\Phi\left(\frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}_1\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{Cx} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - \left(\mathbf{C}\boldsymbol{\gamma}_1\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\mathbf{C}\boldsymbol{\gamma}_1}}\right).$$

If $\boldsymbol{\gamma}_1 \in \mathcal{N}\left(\mathbf{C}\right)$, then the marginal density function reduces to

$$f\left(\mathbf{Cx}|\Pi_i\right) = \varphi_{p-q}\left(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right).$$

Proof: The proof follows from part (ii) of Proposition 2.2.

Lemma 2.10. Let $\mathbf{x} \sim SN_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}\right)$, where $\mathbf{H} \in \mathbb{R}_{p\times p}$ and $rank\left(\mathbf{H}\right) = p$. Then, $p\left(\mathbf{Hx}|\boldsymbol{\theta}\right) = p\left(\mathbf{x}|\boldsymbol{\theta}\right)$, where $p\left(\cdot|\boldsymbol{\theta}\right)$ is the p-dimensional skew-normal density function with parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}\}$.

Proof: Let $p\left(\cdot|\theta\right)$ denote the MSN density with parameters $\boldsymbol{\theta}$. Then,

$$\begin{aligned}
p\left(\mathbf{Hx}|\boldsymbol{\theta}\right) &= \frac{1}{\Phi\left(\delta_0\right)}\varphi_p\left(\mathbf{Hx}; \mathbf{H}\boldsymbol{\mu}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)\Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\mathbf{H}'\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\left(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu}\right)}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{H}'\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\mathbf{H}\boldsymbol{\gamma}}}\right) \\
&= \frac{1}{\Phi\left(\delta_0\right)}\frac{1}{\left(2\pi\right)^{n/2}\left|\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right|^{1/2}}\exp\left\{-\frac{1}{2}\left(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu}\right)'\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\left(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu}\right)\right\} \times \\
&\qquad \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\mathbf{H}\mathbf{H}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{-1}\mathbf{H}(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{H}\mathbf{H}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{-1}\mathbf{H}\boldsymbol{\gamma}}}\right) \\
&= \frac{1}{\Phi\left(\delta_0\right)}\varphi_p\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right)\Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}\right) \\
&= p\left(\mathbf{x}|\boldsymbol{\theta}\right).
\end{aligned}$$

Next, we present a series of theorems that form the main results of this chapter. The theorems utilize the Moore-Penrose pseudoinverse $\mathbf{F}^+$, where $\mathbf{M} = \mathbf{FG}$ of a full-rank decomposition in order to obtain a linear compression matrix that preserves the full-feature Bayes assignment of the vector $\mathbf{x} \in \mathbb{R}_{p\times 1}$ to the MSN population Π_k for $k \in \{1, ..., m\}$.

Theorem 2.2 $\left(\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j, \delta_{0i} = \delta_{0j}\right)$. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by the distribution $SN_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}\right)$ with location parameter $\boldsymbol{\mu}_i$

such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$ for some $k \in \{2, ..., m\}$, skewness parameter $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}$, dispersion parameter $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, and scalar $\delta_{0i} = \delta_0$ for $i \in \{1, 2, ..., m\}$. Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) | \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_3 - \boldsymbol{\mu}_1) | ... | \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1)\right],$$

where $\mathbf{M} = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}$ with $rank(\mathbf{M}) = q < p$, and let $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^{+}\right)$ and $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$. Also, let $\mathbf{x}$ be an unlabeled observation belonging to Π_r for $r \in \{1, ..., m\}$. Then, the p-variate Bayes classifier assigns $\mathbf{x}$ to Π_k for $k \in \{1, ..., m\}$ if and only if the q-dimensional Bayes classifier assigns $\mathbf{F}'\mathbf{x}$ to Π_k.

Proof: Let $\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^{+}_{q\times p} \\ \mathbf{C}_{(p-q)\times p} \end{bmatrix} \mathbf{x}$, where $\mathbf{x} \sim SN_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$ for some $k \in \{1, 2, ..., m\}$. Also, let $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$ with $\mathbf{C} = \mathbf{R}(\mathbf{I} - \mathbf{FF}^{+})$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ is $rank(\mathbf{R}) = p - q$ so that $rank(\mathbf{H}) = p$. Then, $\mathbf{w} \sim SN(\mathbf{H}\boldsymbol{\mu}_k, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}', \delta_0, \mathbf{H}\boldsymbol{\gamma})$, where

$$\mathbf{H}\boldsymbol{\mu}_k = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\mu}_k \\ \mathbf{C}\boldsymbol{\mu}_k \end{bmatrix} \text{ and } \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime} & \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{C}' \\ \mathbf{C}\boldsymbol{\Sigma}\mathbf{F}^{+\prime} & \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' \end{bmatrix}, \; k \in \{1, 2, ..., m\}.$$

By Lemma 2.8,

$$g(\mathbf{u}|\mathbf{y}) = \frac{1}{\Phi(\delta_0)} \varphi_q\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_k, \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_0 + \mathbf{l}_1'\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_k\right)\right),$$

where

$$l_0 = \frac{\delta_0}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{F}^{+\prime}\left(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}\right)\mathbf{F}^{+}\boldsymbol{\gamma}}} \tag{2.13}$$

and

$$\mathbf{l}_1 = \frac{\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-1/2}\mathbf{F}^{+}\boldsymbol{\gamma}}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{F}^{+\prime}\left(\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}\right)\mathbf{F}^{+}\boldsymbol{\gamma}}}. \tag{2.14}$$

Also, from Lemma 2.9, the marginal density of $\mathbf{y}$ is

$$h\left(\mathbf{y}\right)=\varphi_{p-q}\left(\mathbf{Cx};\mathbf{C}\boldsymbol{\mu}_{1},\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)\Phi\left(\frac{\delta_{0}+\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{Cx}-\mathbf{C}\boldsymbol{\mu}_{1}\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right).$$

Now recall that the p-variate Bayes classification procedure assigns $\mathbf{x}$ to Π_j if and only if

$$\alpha_{j}\frac{1}{\Phi\left(\delta_{0}\right)}\varphi\left(\mathbf{x};\boldsymbol{\mu}_{j},\boldsymbol{\Sigma}\right)\Phi\left(\frac{\delta_{0}+\boldsymbol{\gamma}'\ \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_{j}\right)}{\sqrt{1-\boldsymbol{\gamma}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}\right)>$$

$$\alpha_{i}\frac{1}{\Phi\left(\delta_{0}\right)}\varphi\left(\mathbf{x};\boldsymbol{\mu}_{i},\boldsymbol{\Sigma}\right)\Phi\left(\frac{\delta_{0}+\boldsymbol{\gamma}'\ \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_{i}\right)}{\sqrt{1-\boldsymbol{\gamma}'\,\boldsymbol{\Sigma}^{-1}\boldsymbol{\gamma}}}\right)$$

for $i=1,...,m$, $i\neq j$, which is equivalent to

$$\alpha_{j}\varphi\left(\mathbf{Hx};\mathbf{H}\boldsymbol{\mu}_{j},\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)\Phi\left(\frac{\delta_{0}+\boldsymbol{\gamma}'\mathbf{H}'\ \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\left(\mathbf{Hx}-\boldsymbol{H}\boldsymbol{\mu}_{j}\right)}{\sqrt{1-\boldsymbol{\gamma}'\mathbf{H}'\ \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\mathbf{H}\boldsymbol{\gamma}}}\right)>$$

$$\alpha_{i}\varphi\left(\mathbf{Hx};\mathbf{H}\boldsymbol{\mu}_{i},\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)\Phi\left(\frac{\delta_{0}+\boldsymbol{\gamma}'\mathbf{H}'\ \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\left(\mathbf{Hx}-\mathbf{H}\boldsymbol{\mu}_{i}\right)}{\sqrt{1-\boldsymbol{\gamma}'\mathbf{H}'\ \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\mathbf{H}\boldsymbol{\gamma}}}\right)$$

for $i=1,...,m$, $i\neq j$. Hence, for $i,j=1,2,...,m$ and $i\neq j$, we have

$$\alpha_{j}\varphi_{q}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_{j},\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)\Phi\left(l_{0}+\mathbf{l}_{1}'\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_{j}\right)\right)\times$$

$$\varphi_{p-q}\left(\mathbf{Cx};\mathbf{C}\boldsymbol{\mu}_{1},\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)\Phi\left(\frac{\delta_{0}+\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{Cx}-\mathbf{C}\boldsymbol{\mu}_{1}\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)>$$

$$\alpha_{i}\varphi_{q}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_{i},\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)\Phi\left(l_{0}+\mathbf{l}_{1}'\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_{i}\right)\right)\times$$

$$\varphi_{p-q}\left(\mathbf{Cx};\mathbf{C}\boldsymbol{\mu}_{1},\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)\Phi\left(\frac{\delta_{0}+\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{Cx}-\mathbf{C}\boldsymbol{\mu}_{1}\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)$$

for $i=1,...,m$, $i\neq j$ by Lemma 2.8. Finally, because

$$\varphi_{p-q}\left(\mathbf{Cx};\mathbf{C}\boldsymbol{\mu}_{1},\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)\ \text{and}\ \Phi\left(\frac{\delta_{0}+\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{Cx}-\mathbf{C}\boldsymbol{\mu}_{1}\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)$$

do not depend on k for $k=2,...,m$, we have that

$$\alpha_{j}\varphi_{q}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_{j},\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)\Phi\left(l_{0}+\mathbf{l}_{1}'\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_{j}\right)\right)>$$

$$\alpha_i \varphi_q \left(\mathbf{F}^+ \mathbf{x}; \mathbf{F}^+ \boldsymbol{\mu}_i, \mathbf{F}^+ \boldsymbol{\Sigma} \mathbf{F}^{+\prime} \right) \Phi \left(l_0 + \mathbf{l}_1' \left(\mathbf{F}^+ \boldsymbol{\Sigma} \mathbf{F}^{+\prime} \right)^{-\frac{1}{2}} \left(\mathbf{F}^+ \mathbf{x} - \mathbf{F}^+ \boldsymbol{\mu}_i \right) \right),$$

where l_0 and $\mathbf{l}_1$ are defined in (2.13) and (2.14), respectively.

Hence, if the p-variate Bayes classifier assigns the unlabeled observation $\mathbf{x}$ into Π_j, then the q-variate Bayes classifier assigns $\mathbf{F}'\mathbf{x}$ into Π_j because $\mathcal{C}(\mathbf{F}^+) = \mathcal{C}(\mathbf{F}')$. The preceding arguments are reversible and, thus, the original p-variate Bayes classification assignment is preserved by the linear transformation $\mathbf{y} = \mathbf{F}'\mathbf{x}$.

The following corollary provides the conditions in which the linear dimension reduction matrix reduces to Fisher's linear discriminant function.

Corollary 2.1. Let Π_i, $i = 1, 2$, be represented by p-dimensional *MSN* populations with *a priori* probability $\alpha_i > 0$, mean vector $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, skewness parameter $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2$, dispersion parameter $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, and scalar parameter $\delta_0 = \delta_{01} = \delta_{02}$. Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \right],$$

where $rank(\mathbf{M}) = 1 < p$. Then, the p-variate Bayes procedure assigns $\mathbf{x}$ to Π_k for $k = 1, 2$ if and only if the one-dimensional Bayes procedure assigns $\mathbf{M}'\mathbf{x}$ to Π_k.

Proof: The proof consists of substitution of $\mathbf{M}$ for $\mathbf{F}$ and $\mathbf{M}^+$ for $\mathbf{F}^+$ in Theorem 2.2.

Theorem 2.3 $\left(\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j, \delta_{0i} = \delta_{0j} \right)$. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by the distribution $SN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma})$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$, dispersion parameter $\boldsymbol{\Sigma}_i$ such that $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$ for some $i, j \in \{2, ..., m\}$, skewness parameter $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}$, and skew scalar $\delta_{0i} = \delta_0$ for $i \in \{1, 2, ..., m\}$. Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) | ... | \boldsymbol{\Sigma}_m^{-1} (\boldsymbol{\mu}_m - \boldsymbol{\mu}_1) | \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 | ... | \boldsymbol{\Sigma}_m - \boldsymbol{\Sigma}_1 \right],$$

where $\mathbf{M} = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}$ with $rank(\mathbf{M}) = q < p$, and let $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^{+}\right)$ and $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$. Also, let $\mathbf{x}$ be an unlabeled observation belonging to Π_j for $j \in \{1, ..., m\}$. Then, the p-variate Bayes classifier assigns $\mathbf{x}$ to Π_k for $k \in \{1, ..., m\}$ if and only if the q-dimensional Bayes classifier assigns $\mathbf{F}^{+}\mathbf{x}$ to Π_k.

Proof: Let $\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^{+}_{q\times p} \\ \mathbf{C}_{(p-q)\times p} \end{bmatrix} \mathbf{x}$, where $\mathbf{x} \sim SN_p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \delta_0, \boldsymbol{\gamma}\right)$ for $k \in \{1, 2, ..., m\}$. Let $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}^{\prime} \end{bmatrix}^{\prime}$ with $\mathbf{C} = \mathbf{R}(\mathbf{I} - \mathbf{FF}^{+})$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ with $rank(\mathbf{R}) = p - q$ so that $rank\left(\mathbf{H}\right) = p$. Then, $\mathbf{w} \sim SN\left(\mathbf{H}\boldsymbol{\mu}_k, \mathbf{H}\boldsymbol{\Sigma}_k\mathbf{H}^{\prime}, \delta_0, \mathbf{H}\boldsymbol{\gamma}\right)$, where

$$\mathbf{H}\boldsymbol{\mu}_k = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\mu}_k \\ \mathbf{C}\boldsymbol{\mu}_k \end{bmatrix} \text{ and } \mathbf{H}\boldsymbol{\Sigma}_k\mathbf{H}^{\prime} = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime} & \mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{C}^{\prime} \\ \mathbf{C}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime} & \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}^{\prime} \end{bmatrix}, \; k \in \{1, 2, ..., m\}.$$

By Lemma 2.8,

$$g\left(\mathbf{u}|\mathbf{y}\right) = \frac{1}{\Phi\left(\delta_0\right)}\varphi_q\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_k, \mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime}\right)\Phi\left(l_{0k} + \mathbf{l}_{1k}^{\prime}\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_k\right)\right),$$

where

$$l_{0k} = \frac{\delta_0}{\sqrt{1 - \boldsymbol{\gamma}^{\prime}\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\boldsymbol{\Sigma}_k^{-1}\mathbf{F}\right)\mathbf{F}^{+}\boldsymbol{\gamma}}} \tag{2.15}$$

and

$$\mathbf{l}_{1k} = \frac{\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime}\right)^{-1/2}\mathbf{F}^{+}\boldsymbol{\gamma}}{\sqrt{1 - \boldsymbol{\gamma}^{\prime}\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\boldsymbol{\Sigma}_k^{-1}\mathbf{F}\right)\mathbf{F}^{+}\boldsymbol{\gamma}}}. \tag{2.16}$$

Also, by Lemma 2.9, the marginal density of $\mathbf{y}$ is

$$h\left(\mathbf{y}\right) = \varphi_{p-q}\left(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}^{\prime}\right)\Phi\left(\frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}^{\prime}\right)^{-1}\left(\mathbf{Cx} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - \left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}^{\prime}\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right).$$

The p-variate Bayes procedure assigns $\mathbf{x}$ to Π_j if and only if

$$\alpha_j \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_j\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\gamma}}}\right) >$$

$$\alpha_i \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\gamma}}}\right)$$

for $i = 1, ..., m$, $i \neq j$, which, by Lemma 2.10, implies

$$\alpha_j \varphi\left(\mathbf{H}\mathbf{x}; \mathbf{H}\boldsymbol{\mu}_j, \mathbf{H}\boldsymbol{\Sigma}_j\mathbf{H}'\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_j\mathbf{H}'\right)^{-1}\left(\mathbf{H}\mathbf{x} - \boldsymbol{H}\boldsymbol{\mu}_j\right)}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_j\mathbf{H}'\right)^{-1}\mathbf{H}\boldsymbol{\gamma}}}\right) >$$

$$\alpha_i \varphi\left(\mathbf{H}\mathbf{x}; \mathbf{H}\boldsymbol{\mu}_i, \mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}'\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}'\right)^{-1}\left(\mathbf{H}\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_i\right)}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}'\right)^{-1}\mathbf{H}\boldsymbol{\gamma}}}\right)$$

for $i = 1, ..., m$, $i \neq j$. Hence, for $i, j = 1, 2, ..., m$ and $i \neq j$, we have

$$\alpha_j \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_j, \mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}'_{1j}\left(\mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_j\right)\right) \times$$

$$\varphi_{p-q}\left(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right) \Phi\left(\frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - \left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right) >$$

$$\alpha_i \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}'_{1i}\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right) \times$$

$$\varphi_{p-q}\left(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right) \Phi\left(\frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - \left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)$$

for $i = 1, ..., m$, $i \neq j$ by Lemma 2.8. Finally, because

$$\varphi_{p-q}\left(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right) \Phi\left(\frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - \left(\mathbf{C}\gamma\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{-1}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)$$

does not depend on k for $k = 2, ..., m$, we have

$$\alpha_j \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_j, \mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}'_{1j}\left(\mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_j\right)\right) >$$

$$\alpha_i \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}'_{1i}\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right),$$

where l_{0k} and $\mathbf{l}_{1k}$ are defined for Π_k in (2.15) and (2.16), respectively. Thus, if the unlabeled observation $\mathbf{x}$ is classified into Π_j, then $\mathbf{F}^{+}\mathbf{x}$ is classified into Π_j. The preceding arguments are reversible and, therefore, the original p-variate Bayes classification assignment is preserved by the linear transformation $\mathbf{y} = \mathbf{F}^{+}\mathbf{x}$.

Theorem 2.4. $\left(\boldsymbol{\gamma}_i \neq \boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j, \delta_{0i} = \delta_{0j}\right)$. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by the distribution $SN_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}_i\right)$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$ and skewness parameter $\boldsymbol{\gamma}_i$ such that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_k$ for some $k \in \{2, ..., m\}$, dispersion parameter $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, and scalar $\delta_{0i} = \delta_0$ for $i \in \{1, 2, ..., m\}$. Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right) | ... | \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1\right) | \boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1 | ... | \boldsymbol{\gamma}_m - \boldsymbol{\gamma}_1\right],$$

where $\mathbf{M} = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}$ with $rank(\mathbf{M}) = q < p$, and let $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$, where $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{F}\mathbf{F}^{+}\right)$ with $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$. Also, let $\mathbf{x}$ be an unlabeled observation belonging to Π_j for $j \in \{1, ..., m\}$. Then, the p-variate Bayes procedure assigns $\mathbf{x}$ to Π_k for $k \in \{1, ..., m\}$ if and only if the q-dimensional Bayes procedure assigns $\mathbf{F}^{+}\mathbf{x}$ to Π_k.

Proof: Let $\mathbf{w} = \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^{+}_{q\times p} \\ \mathbf{C}_{(p-q)\times p} \end{bmatrix} \mathbf{x}$, where $\mathbf{x} \sim SN_p\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}_k\right)$ for some $k \in \{1, 2, ..., m\}$. Let $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$ with $\mathbf{C} = \mathbf{R}(\mathbf{I} - \mathbf{F}\mathbf{F}^{+})$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{R}) = p - q$, and, thus, $rank\left(\mathbf{H}\right) = p$. Then, $\mathbf{w} \sim SN\left(\mathbf{H}\boldsymbol{\mu}_k, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}', \delta_0, \mathbf{H}\boldsymbol{\gamma}_k\right)$, where

$$\mathbf{H}\boldsymbol{\mu}_k = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\mu}_k \\ \mathbf{C}\boldsymbol{\mu}_k \end{bmatrix} \text{ and } \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime} & \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{C}' \\ \mathbf{C}\boldsymbol{\Sigma}\mathbf{F}^{+\prime} & \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' \end{bmatrix}, \; k \in \{1, 2, ..., m\}.$$

By Lemma 2.8,

$$g\left(\mathbf{u}|\mathbf{y}\right) = \frac{1}{\Phi\left(\delta_0\right)} \varphi_q\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_k, \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_{0k} + \mathbf{l}'_{1k}\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_k\right)\right),$$

where

$$l_{0k} = \frac{\delta_0}{\sqrt{1 - \boldsymbol{\gamma}_k' \mathbf{F}^{+\prime} \left(\mathbf{F}' \boldsymbol{\Sigma}^{-1} \mathbf{F}\right) \mathbf{F}^{+} \boldsymbol{\gamma}_k}} \tag{2.17}$$

and

$$\mathbf{l}_{1k} = \frac{\left(\mathbf{F}^{+} \boldsymbol{\Sigma} \mathbf{F}^{+\prime}\right)^{-1/2} \mathbf{F}^{+} \boldsymbol{\gamma}_k}{\sqrt{1 - \boldsymbol{\gamma}_k' \mathbf{F}^{+\prime} \left(\mathbf{F}' \boldsymbol{\Sigma}^{-1} \mathbf{F}\right) \mathbf{F}^{+} \boldsymbol{\gamma}_k}}. \tag{2.18}$$

Also, by Lemma 2.9 and the fact that $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$, the marginal density of $\mathbf{y}$ is

$$h(\mathbf{y}) = \varphi_{p-q}(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}').$$

Let $p(\cdot|\Pi_i)$ denote the p-dimensional MSN density corresponding to population Π_i, $i = 1, ..., m$. Recall that the p-variate Bayes procedure assigns $\mathbf{x}$ to Π_j if and only if

$$\alpha_j \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_j' \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_j\right)}{\sqrt{1 - \boldsymbol{\gamma}_j' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}_j}}\right) >$$

$$\alpha_i \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_i' \boldsymbol{\Sigma}^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)}{\sqrt{1 - \boldsymbol{\gamma}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\gamma}_i}}\right)$$

for $i = 1, ..., m$, $i \neq j$, which is equivalent to

$$\alpha_j \varphi\left(\mathbf{H}\mathbf{x}; \mathbf{H}\boldsymbol{\mu}_j, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_j' \mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\left(\mathbf{H}\mathbf{x} - \boldsymbol{H}\boldsymbol{\mu}_j\right)}{\sqrt{1 - \boldsymbol{\gamma}_j' \mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1} \mathbf{H}\boldsymbol{\gamma}_j}}\right) >$$

$$\alpha_i \varphi\left(\mathbf{H}\mathbf{x}; \mathbf{H}\boldsymbol{\mu}_i, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_i' \mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1}\left(\mathbf{H}\mathbf{x} - \mathbf{H}\boldsymbol{\mu}_i\right)}{\sqrt{1 - \boldsymbol{\gamma}_i' \mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\right)^{-1} \mathbf{H}\boldsymbol{\gamma}_i}}\right)$$

for $i = 1, ..., m$, $i \neq j$. Hence, for $i, j = 1, 2, ..., m$ and $i \neq j$, we have

$$\alpha_j \varphi_q\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_j, \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}_{1j}' \left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_j\right)\right) \times$$

$$\varphi_{p-q}(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}') >$$

$$\alpha_i \varphi_q\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_i, \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}_{1i}' \left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_i\right)\right) \times$$

$$\varphi_{p-q}(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$$

for $i = 1, ..., m$, $i \neq j$ by Lemma 2.8. Finally, we have that

$$\alpha_j \varphi_q\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_j, \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}_{1j}' \left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_j\right)\right) >$$

$$\alpha_i \varphi_q \left(\mathbf{F}^+ \mathbf{x}; \mathbf{F}^+ \boldsymbol{\mu}_i, \mathbf{F}^+ \boldsymbol{\Sigma} \mathbf{F}^{+\prime} \right) \Phi \left(l_{0i} + \mathbf{l}'_{1i} \left(\mathbf{F}^+ \boldsymbol{\Sigma} \mathbf{F}^{+\prime} \right)^{-\frac{1}{2}} \left(\mathbf{F}^+ \mathbf{x} - \mathbf{F}^+ \boldsymbol{\mu}_i \right) \right),$$

where l_{0k} and $\mathbf{l}_{1k}$ are defined for the k^{th} class in (2.17) and (2.18), respectively, because $\varphi_{p-q} \left(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' \right)$ does not depend on k for $k = 2, ..., m$. Hence, if the p-variate Bayes classifier assigns the unlabeled observation $\mathbf{x}$ into Π_j, then the q-variate Bayes classifier assigns $\mathbf{F}^+ \mathbf{x}$ into Π_j. The preceding arguments are reversible and, thus, the original p-variate Bayes classification assignment is preserved by the linear transformation $\mathbf{y} = \mathbf{F}^+ \mathbf{x}$.

Corollary 2.2. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by distribution $SN_p \left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}_i \right)$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$, skew parameter $\boldsymbol{\gamma}_i$ such that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_k$ for some $k \in \{2, ..., m\}$, dispersion parameter $\boldsymbol{\Sigma}$, and scalar δ_0 for $i \in \{2, ..., m\}$. In addition, let Π_1 be represented by the multivariate distribution $N_p \left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma} \right)$. Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \right) | ... | \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1 \right) | \boldsymbol{\gamma}_2 | ... | \boldsymbol{\gamma}_m \right],$$

where $\mathbf{M} = \mathbf{F}\mathbf{G}$ is a full-rank decomposition of $\mathbf{M}$ with $rank \left(\mathbf{M} \right) = q < p$, and let $\boldsymbol{\gamma}_1 \in \mathcal{N} \left(\mathbf{C} \right)$, where $\mathbf{C} = \mathbf{R} \left(\mathbf{I} - \mathbf{F}\mathbf{F}^+ \right)$ and $\mathbf{R} \in \mathbb{R}_{(p-q) \times p}$. Also, let $\mathbf{x}$ be an unlabeled observation vector belonging to Π_j for $j \in \{1, 2, ..., m\}$. Then, the p-variate Bayes classifier assigns $\mathbf{x}$ to Π_k for $k \in \{1, 2, ..., m\}$ if and only if the q-dimensional Bayes classifier assigns $\mathbf{F}^+ \mathbf{x}$ to Π_k.

Proof: The proof follows immediately because $\boldsymbol{\gamma}_1 = \mathbf{0}$.

Theorem 2.5 $\left(\boldsymbol{\gamma}_i \neq \boldsymbol{\gamma}_j, \boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j, \delta_{0i} = \delta_{0j} \right)$ for $i = 1, 2, ..., m$. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by distribution $SN_p \left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i \right)$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$, skewness parameter $\boldsymbol{\gamma}_i$ such that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_k$, dispersion parameter $\boldsymbol{\Sigma}_i$ such that $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_k$, and scalar δ_0 for some $k \in \{2, ..., m\}$.

Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1\right)|...|\boldsymbol{\Sigma}_m^{-1}\left(\boldsymbol{\mu}_m-\boldsymbol{\mu}_1\right)|\boldsymbol{\Sigma}_2-\boldsymbol{\Sigma}_1|...|\boldsymbol{\Sigma}_m-\boldsymbol{\Sigma}_1|\boldsymbol{\gamma}_2-\boldsymbol{\gamma}_1|...|\boldsymbol{\gamma}_m-\boldsymbol{\gamma}_1\right],$$

where $\mathbf{M}=\mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}$ with $rank(\mathbf{M})=q<p$, and let $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$. Also, let $\mathbf{x}$ be an unlabeled observation vector belonging to Π_k for $j \in \{1,2,...,m\}$. Then, the p-variate Bayes classifier assigns $\mathbf{x}$ to Π_k for $k \in \{1,...,m\}$ if and only if the q-dimensional Bayes classifier assigns $\mathbf{F}^{+}\mathbf{x}$ to Π_k.

Proof: Let $\mathbf{w}=\begin{bmatrix}\mathbf{u}\\ \mathbf{y}\end{bmatrix}=\begin{bmatrix}\mathbf{F}^{+}_{q\times p}\\ \mathbf{C}_{(p-q)\times p}\end{bmatrix}\mathbf{x}$, where $\mathbf{x}\sim SN_p\left(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\delta_0,\boldsymbol{\gamma}_k\right)$, $k=1,2,...,m$. Let $\mathbf{H}$ be a full-rank $p\times p$ matrix defined by $\mathbf{H}\equiv\begin{bmatrix}\mathbf{F}^{+}\\ \mathbf{C}\end{bmatrix}$ with $\mathbf{C}=\mathbf{R}(\mathbf{I}-\mathbf{F}\mathbf{F}^{+})$ with $\mathbf{R}\in\mathbb{R}_{(p-q)\times p}$ full rank such that $rank(\mathbf{R})=p-q$. Then, $\mathbf{w}\sim SN\left(\mathbf{H}\boldsymbol{\mu}_k,\mathbf{H}\boldsymbol{\Sigma}_k\mathbf{H}',\delta_0,\mathbf{H}\boldsymbol{\gamma}_k\right)$, where

$$\mathbf{H}\boldsymbol{\mu}_k=\begin{bmatrix}\mathbf{F}^{+}\boldsymbol{\mu}_k\\ \mathbf{C}\boldsymbol{\mu}_k\end{bmatrix} \text{ and } \mathbf{H}\boldsymbol{\Sigma}_k\mathbf{H}'=\begin{bmatrix}\mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime} & \mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{C}'\\ \mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} & \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\end{bmatrix},$$

for $i=1,2,...,m$. By Lemma 2.8,

$$g\left(\mathbf{u}|\mathbf{y}\right)=\frac{1}{\Phi\left(\delta_0\right)}\varphi_q\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_k,\mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime}\right)\Phi\left(l_{0k}+\mathbf{l}_{1k}'\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_k\right)\right),$$

where

$$l_{0k}=\frac{\delta_0}{\sqrt{1-\boldsymbol{\gamma}_k'\mathbf{F}^{+\prime}\left(\mathbf{F}'\boldsymbol{\Sigma}_k^{-1}\mathbf{F}\right)\mathbf{F}^{+}\boldsymbol{\gamma}_k}} \tag{2.19}$$

and

$$\mathbf{l}_{1k}=\frac{\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_k\mathbf{F}^{+\prime}\right)^{-1/2}\mathbf{F}^{+}\boldsymbol{\gamma}_k}{\sqrt{1-\boldsymbol{\gamma}_k'\mathbf{F}^{+\prime}\left(\mathbf{F}'\boldsymbol{\Sigma}_k^{-1}\mathbf{F}\right)\mathbf{F}^{+}\boldsymbol{\gamma}_k}}. \tag{2.20}$$

Also, by Lemma 2.9 and the fact that $\boldsymbol{\gamma}_1\in\mathcal{N}(\mathbf{C})$, the marginal density of $\mathbf{y}$ is

$$h\left(\mathbf{y}|\Pi_k\right)=\varphi_{p-q}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_1,\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right). \tag{2.21}$$

The p-variate Bayes classification procedure assigns an unlabeled observation $\mathbf{x}$ to Π_j if and only if

$$\alpha_j \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_j' \boldsymbol{\Sigma}_j^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_j\right)}{\sqrt{1 - \boldsymbol{\gamma}_j' \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\gamma}_j}}\right) >$$

$$\alpha_i \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_i' \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_i\right)}{\sqrt{1 - \boldsymbol{\gamma}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\gamma}_i}}\right)$$

for $i = 1, 2, ..., m$, $i \neq j$, which, by Lemma 2.10, implies

$$\alpha_j \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{Hx}; \mathbf{H}\boldsymbol{\mu}_j, \mathbf{H}\boldsymbol{\Sigma}_j\mathbf{H}'\right) \Phi\left(\frac{\delta_0 + \mathbf{H}\boldsymbol{\gamma}_j'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_j\mathbf{H}'\right)^{-1}\left(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu}_j\right)}{\sqrt{1 - \boldsymbol{\gamma}_j'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_j\mathbf{H}'\right)^{-1} \mathbf{H}\boldsymbol{\gamma}_j}}\right) >$$

$$\alpha_i \frac{1}{\Phi(\delta_0)} \varphi\left(\mathbf{Hx}; \mathbf{H}\boldsymbol{\mu}_i, \mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}'\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_i'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}'\right)^{-1}\left(\mathbf{Hx} - \boldsymbol{H}\boldsymbol{\mu}_i\right)}{\sqrt{1 - \boldsymbol{\gamma}_i'\mathbf{H}' \left(\mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}'\right)^{-1} \mathbf{H}\boldsymbol{\gamma}_i}}\right).$$

Hence, for $i, j = 1, 2, ..., m$ and $i \neq j$, we have

$$\alpha_j \frac{1}{\Phi(\beta_{0j})} \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_j, \mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}_{1j}' \left(\mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_j\right)\right) \times$$

$$\varphi_{p-q}\left(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right) >$$

$$\alpha_i \frac{1}{\Phi(\beta_{0i})} \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}_{1i}' \left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right) \times$$

$$\varphi_{p-q}\left(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)$$

for $i = 1, 2, ..., m$, $i \neq j$ by Lemma 2.8. Therefore,

$$\alpha_j \frac{1}{\Phi(\beta_{0j})} \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_j, \mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}_{1j}' \left(\mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_j\right)\right) >$$

$$\alpha_i \frac{1}{\Phi(\beta_{0i})} \varphi_q\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}_{1i}' \left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{-\frac{1}{2}} \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right),$$

because $\varphi_{p-q}\left(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)$ does not depend on k for $k = 2, ..., m$ and because $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$, where l_{0k} and $\mathbf{l}_{1k}$ are defined in (2.19) and (2.20) for Π_k, respectively. Hence, if p-variate Bayes classifier assigns the unlabeled observation $\mathbf{x}$ into Π_j, then the q-variate Bayes classifier assigns $\mathbf{F}^+\mathbf{x}$ into Π_j. The preceding arguments are

reversible, and, thus, the original p-variate Bayes classification is preserved by the linear transformation $\mathbf{y} = \mathbf{F}^{+}\mathbf{x}$.

Corollary 2.3. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by the distribution $SN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$, dispersion parameter $\boldsymbol{\Sigma}_i$ such that $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_k$, skew parameter $\boldsymbol{\gamma}_i$ such that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_k$ for some $k \in \{2, ..., m\}$, and scalar δ_0 for $i \in \{2, ..., m\}$. In addition, let Π_1 be represented by the distribution $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Next, let

$$\mathbf{M} \equiv \left[\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) | ... | \boldsymbol{\Sigma}_m^{-1}(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1) | \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 | ... | \boldsymbol{\Sigma}_m - \boldsymbol{\Sigma}_1 | \boldsymbol{\gamma}_2 | ... | \boldsymbol{\gamma}_m\right],$$

where $\mathbf{M} = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}$ with $rank(\mathbf{M}) = q < p$, and let $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$, where $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^{+}\right)$ and $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$. Also, let $\mathbf{x}$ be an unlabeled observation vector belonging to Π_j for $j \in \{1, 2, ..., m\}$. Then, the p-variate Bayes classification procedure assigns $\mathbf{x}$ to Π_k for $k \in \{1, 2, ..., m\}$ if and only if the q-dimensional Bayes classification procedure assigns $\mathbf{F}^{+}\mathbf{x}$ to Π_k.

Proof: The proof follows immediately from Theorem 2.5 because $\boldsymbol{\gamma}_1 = \mathbf{0}$.

If $rank(\mathbf{M}) = p$, one can use Theorems 2.2 through 2.5 to obtain a $q \times p$ linear compression matrix that preserves the full-feature PMC. Also, in many situations when Theorems 2.2 through 2.5 hold, we may desire to determine a low-dimensional representation with dimension less than q, say dimension r, where $1 \leq r < q < p$. Thus, we seek to construct an r-dimensional representation of the density $p(\mathbf{x}|\Pi_i)$, $i = 1, 2, ..., m$, which preserves the original p-dimensional $BPMC$ as much as possible.

One method of approximating $\mathbf{M}$ by $rank(r)$ matrix, where $1 \leq r < p$, is the *singular value decomposition* (SVD) approximation to $\mathbf{M}$. We use the following theorem to determine a $rank(r)$ approximation to the linear sufficient matrix $\mathbf{M}$.

Theorem 2.6. (Tubbs et al. (1982)): Let $\mathbf{C}_{s,t}^{(p)}$ denote the class of all $s \times t$ real ma-

trices of rank p, and let $C^{(r)}$ denote the class of all $s \times t$ real matrices of rank r where $1 \leq r < p$. If $\mathbf{A}_p \in \mathbf{C}_{s,t}^{(p)}$ and $\mathbf{A}_r \in \mathbf{C}_{s,t}^{(r)}$, given by $\mathbf{A}_r = \mathbf{U}\mathbf{D}_r\mathbf{V}'$, then $\|\mathbf{A}_p - \mathbf{A}_r\| < \|\mathbf{A}_p - \mathbf{X}\|$ for all $\mathbf{X} \in \mathbf{C}_{s,t}^{(r)}$, where $\mathbf{A}_p = \mathbf{U}\mathbf{D}_p\mathbf{V}'$, $\mathbf{D}_p = diag\,(d_1, d_2, ..., d_p)$, $\mathbf{D}_r = diag\,(d_1, d_2, ..., d_r, 0_{r+1}, ..., 0_p)$, and $\|\mathbf{A}_k\|$ is the usual Euclidean or Frobenius norm of a matrix $\mathbf{A}_p$, given by $\|\mathbf{A}_p\| = \left(\sum_{i=1}^{s}\sum_{j=1}^{t}|a_{ij}|^2\right) = \left(\sum_{i=1}^{p} d_i^2\right)$. Furthermore, $\|\mathbf{A}_p - \mathbf{A}_r\| = \left(\sum_{i=r+1}^{p} d_i^2\right)^{1/2}$.

2.5 Examples

Example 2.1. In the first example, we demonstrate a low-dimensional representation for three MSN populations having unequal means but equal skew, dispersion, and scalar parameters. Consider the three seven-dimensional populations $SN_7\,(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, $SN_7\,(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, and $SN_7\,(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, where

$$\Pi_1 : \boldsymbol{\mu}_1 = [.25, .5, .75, 1, 1.25, 1.5, 1.75]',$$
$$\Pi_2 : \boldsymbol{\mu}_2 = [.5, 1, 1.5, 2, 2.5, 3, 3.5]', \text{ and}$$
$$\Pi_3 : \boldsymbol{\mu}_3 = [1, 2, 3, 4, 5, 6, 7]'.$$

In addition, we have the common dispersion parameter

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.67 & .15 & .15 & .15 & .15 & .15 & .15 \\ .15 & 4.67 & .15 & .15 & .15 & .15 & .15 \\ .15 & .15 & 4.67 & .15 & .15 & .15 & .15 \\ .15 & .15 & .15 & 4.67 & .15 & .15 & .15 \\ .15 & .15 & .15 & .15 & 4.67 & .15 & .15 \\ .15 & .15 & .15 & .15 & .15 & 4.67 & .15 \\ .15 & .15 & .15 & .15 & .15 & .15 & 4.67 \end{bmatrix}.$$

Also, we have common skew parameter

$$\boldsymbol{\gamma} = [.75, .76, .77, .78, .79, .80, .81]'$$

with $\delta_{0i} = 0$, $i = 1, 2, 3$. Using Theorem 2.2 to formulate a low-dimensional matrix, we obtain

$$\mathbf{M}' = \begin{bmatrix} .014 & .069 & .124 & .180 & .235 & .290 & .345 \\ .041 & .206 & .372 & .539 & .705 & .870 & 1.04 \end{bmatrix}',$$

where $rank\,(\mathbf{M}) = 1$. Therefore, by Theorem 2.2, the three original seven-dimensional density functions can be transformed to the reduced dimension $q = 1$ without increasing the *BPMC*. An approximate one-dimensional representation space is $\mathcal{C}\,(\mathbf{F})$, where

$$\mathbf{F}' = \begin{bmatrix} .024 & .124 & .223 & .322 & .421 & .520 & .619 \end{bmatrix}.$$

Our reduced location parameters are

$$\mu_1 = 2.95, \mu_2 = 5.89, \text{ and } \mu_3 = 11.79$$

with common dispersion parameter $\sigma = 5.28$ and common skew parameter $\gamma = 1.78$.

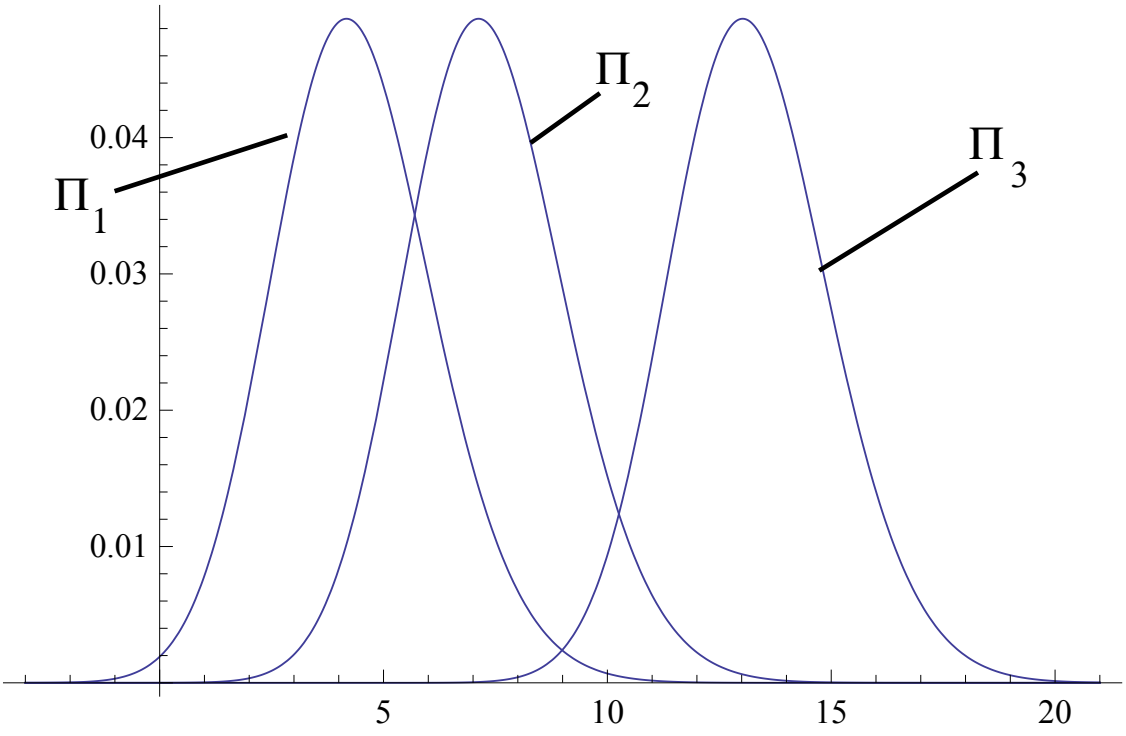

Figure 2.1: The optimal one-dimensional representation for the three seven-dimensional skew-normal densities defined in Example 2.1.

Example 2.2. In the second example, we demonstrate a low-dimensional representation for three *MSN* populations having unequal location, dispersion, and skew parameters. Consider the three six-dimensional populations, $SN_6(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \delta_{01}, \boldsymbol{\gamma}_1)$, $SN_6(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \delta_{02}, \boldsymbol{\gamma}_2)$, and $SN_6(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, \delta_{03}, \boldsymbol{\gamma}_3)$, with

$$\Pi_1 : \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & .2 & .4 & .4 & .1 & .2 \\ .2 & 2 & .4 & .4 & .4 & .4 \\ .4 & .4 & 2 & .2 & .4 & 0 \\ .4 & .4 & .2 & 2 & .4 & .4 \\ .1 & .4 & .4 & .4 & 2 & .4 \\ .2 & .4 & 0 & .4 & .4 & 2 \end{bmatrix}, \boldsymbol{\gamma}_1 = \begin{bmatrix} .763 \\ .698 \\ .698 \\ .821 \\ .829 \\ .745 \end{bmatrix}$$

$$\Pi_2 : \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 4 \\ 4 \\ 2 \\ 2 \\ 3 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.6 & .8 & 1 & 1 & .7 & .8 \\ .8 & 2.6 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2.6 & .8 & 1 & .6 \\ 1 & 1 & .8 & 2.6 & 1 & 1 \\ .7 & 1 & 1 & 1 & 2.6 & 1 \\ .8 & 1 & .6 & 1 & 1 & 2.6 \end{bmatrix}, \boldsymbol{\gamma}_2 = \begin{bmatrix} .963 \\ .898 \\ .898 \\ 1.021 \\ 1.029 \\ .945 \end{bmatrix}$$

$$\Pi_3 : \boldsymbol{\mu}_3 = \begin{bmatrix} 3.93 \\ 4.05 \\ 3.95 \\ 4.05 \\ 4.03 \\ 3.95 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.4 & -.4 & -.2 & -.2 & -.5 & -.4 \\ -.4 & 1.4 & -.2 & -.2 & -.2 & -.2 \\ -.2 & -.2 & 1.4 & -.4 & -.2 & -.6 \\ -.2 & -.2 & -.4 & 1.4 & -.2 & -.2 \\ -.5 & -.2 & -.2 & -.2 & 1.4 & -.2 \\ -.4 & -.2 & -.6 & -.2 & -.2 & 1.4 \end{bmatrix}, \boldsymbol{\gamma}_3 = \begin{bmatrix} 1.063 \\ .998 \\ .998 \\ 1.121 \\ 1.129 \\ 1.045 \end{bmatrix},$$

with $\delta_{0i} = 0$, $i = 1, 2, 3$. As an aside we remark that $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$. Using Theorem 2.5 to formulate a linear dimension reduction matrix, we obtain

$$\mathbf{M} = \begin{bmatrix} .271 & .5 & .6 & .6 & .6 & .6 & .6 & .6 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & .2 & .3 \\ .843 & .5 & .6 & .6 & .6 & .6 & .6 & .6 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & .2 & .3 \\ .877 & .5 & .6 & .6 & .6 & .6 & .6 & .6 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & .2 & .3 \\ -.331 & .5 & .6 & .6 & .6 & .6 & .6 & .6 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & .2 & .3 \\ -.390 & .5 & .6 & .6 & .6 & .6 & .6 & .6 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & .2 & .3 \\ .437 & .5 & .6 & .6 & .6 & .6 & .6 & .6 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & 1.4 & .2 & .3 \end{bmatrix},$$

where $rank\,(\mathbf{M}) = 2$. Therefore, by Theorem 2.5, the original six-dimensional MSN density functions can be transformed into two-dimensional density functions without increasing the probability of misclassification under the Bayes classification assignments. Because $\mathcal{C}\,(\mathbf{F}^{+}) = \mathcal{C}\,(\mathbf{F}')$, an optimal two-dimensional representation space is $\mathcal{C}\,(\mathbf{F})$, where

$$\mathbf{F}' = \begin{bmatrix} .408 & .413 & .413 & .403 & .403 & .409 \\ -.015 & .449 & .476 & -.503 & -.551 & .120 \end{bmatrix}.$$

For Π_1, we now have the reduced parameters

$$\boldsymbol{\mu}_{1R} = \begin{bmatrix} 2.449 \\ -.015 \end{bmatrix}, \boldsymbol{\Sigma}_{1R} = \begin{bmatrix} 3.565 & -.094 \\ -.094 & 1.649 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{1R} = \begin{bmatrix} 1.858 \\ -.147 \end{bmatrix}.$$

For Π_2, we have reduced parameters

$$\boldsymbol{\mu}_{2R} = \begin{bmatrix} 7.368 \\ 1.904 \end{bmatrix}, \boldsymbol{\Sigma}_{2R} = \begin{bmatrix} 7.165 & -.131 \\ -.131 & 1.649 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{2R} = \begin{bmatrix} 2.348 \\ -.152 \end{bmatrix}.$$

The reduced parameters for Π_3 are

$$\boldsymbol{\mu}_{3R} = \begin{bmatrix} 9.777 \\ -.146 \end{bmatrix}, \boldsymbol{\Sigma}_{3R} = \begin{bmatrix} 11.964 & -.180 \\ -.180 & 1.65 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{3R} = \begin{bmatrix} 2.593 \\ -.155 \end{bmatrix}.$$

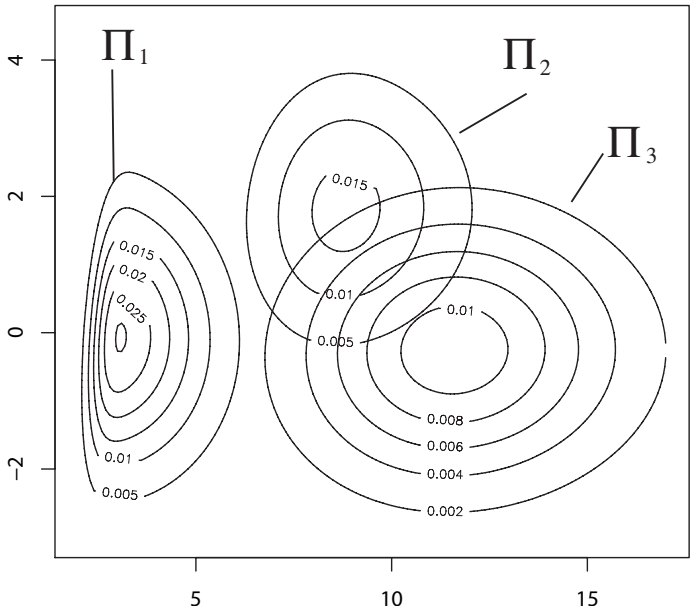

Figure 2.2: The optimal two-dimensional representation for the three six-dimensional skew-normal populations given in Example 2.2.

Immediately above, we portray a two-dimensional representation of the original six-dimensional MSN populations. A one-dimensional representation of these three populations is obtainable if we use the SVD in Theorem 2.6; however, we lose some discriminatory information by applying this procedure. For the first population, we obtain parameters

$$\Pi_1 : \mu_{1R} = 2.449, \sigma_{1R} = 3.57, \text{ and } \gamma_{1R} = 1.86.$$

Also, for the second population, we obtain parameters

$$\Pi_2 : \mu_{2R} = 7.368, \sigma_{2R} = 7.17, \text{ and } \gamma_{2R} = 2.35.$$

Finally, for the third population, we obtain parameters

$$\Pi_3 : \mu_{3R} = 9.777, \sigma_{3R} = 11.96, \text{ and } \gamma_{3R} = 2.59.$$

Example 2.3. In the third example, we use Theorem 2.5 to formulate a low-dimensional representation for three populations with unequal location, dispersion,

and skew parameters with original dimension $p = 7$. Consider the configuration $SN_7(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \delta_{01}, \boldsymbol{\gamma}_1)$, $SN_7(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \delta_{02}, \boldsymbol{\gamma}_2)$, and $SN_7(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, \delta_{03}, \boldsymbol{\gamma}_3)$ with

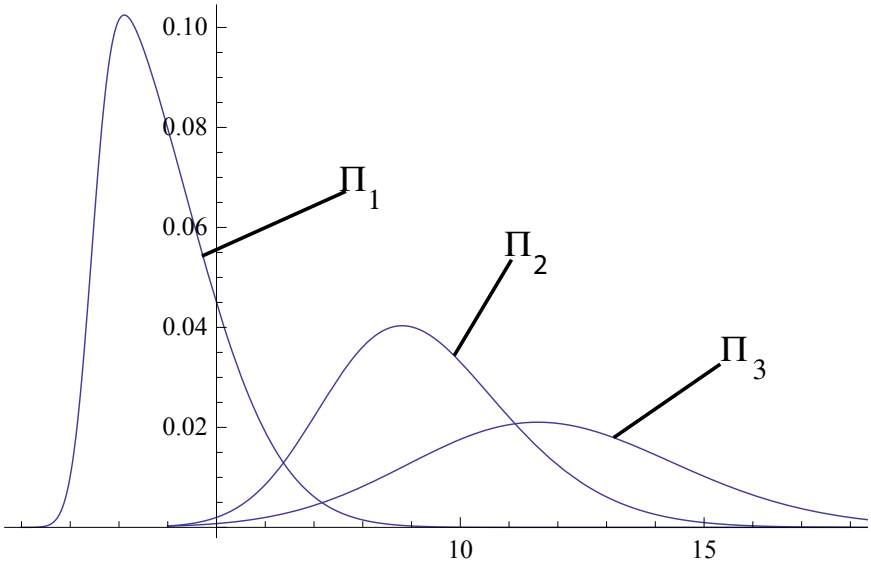

Figure 2.3: A one-dimensional approximate representation for the three six-dimensional skew-normal populations in Example 2.2.

$$\Pi_1 : \boldsymbol{\mu}_1 = [1, 1.5, 3, 4, 2.5, 2, 6]',$$
$$\boldsymbol{\Sigma}_1 = (.10)\,\mathbf{J} + 2.99\mathbf{I}_7,$$
$$\boldsymbol{\gamma}_1 = [.675, .683, .707, .723, .698, .678, .755]'$$

$$\Pi_2 : \boldsymbol{\mu}_2 = [2, 3, 6, 8, 5, 4, 12]',$$
$$\boldsymbol{\Sigma}_2 = (.01)\,\mathbf{J} + 2.99\mathbf{I}_7,$$
$$\boldsymbol{\gamma}_2 = [.575, .583, .607, .623, .598, .655]'$$

$$\Pi_3 : \boldsymbol{\mu}_3 = [3.997, 4.497, 5.997, 6.997, 5.497, 4.997, 8.997]',$$
$$\boldsymbol{\Sigma}_3 = \mathbf{J} + 2.99\mathbf{I}_7,$$
$$\boldsymbol{\gamma}_3 = [1.075, 1.083, 1.107, 1.123, 1.098, 1.078, 1.155]',$$

where $\mathbf{J}$ denotes the 7×7 matrix of 1's and $\mathbf{I}_7$ denotes the 7×7 identity matrix with

$\delta_{0i} = 0$, $i = 1, 2, 3$. Using Theorem 2.5 to formulate a dimension reduction matrix, we obtain

$$\mathbf{M} = \begin{bmatrix} 0.31 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \\ 0.48 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \\ 0.98 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \\ 1.31 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \\ 0.81 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \\ 0.64 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \\ 1.98 & .3 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & -.09 & .9 & .9 & .9 & .9 & .9 & .9 & .9 & -.1 & .4 \end{bmatrix}$$

with $rank\,(\mathbf{M}) = 2$. Therefore, by Theorem 2.5, the original seven-dimensional MSN density functions can be transformed into two-dimensional density functions without increasing the probability of misclassification under the Bayes classification assignments. Because $\mathcal{C}\,(\mathbf{F}^{+}) = \mathcal{C}\,(\mathbf{F}')$, an optimal two-dimensional representation space is $\mathcal{C}\,(\mathbf{F})$, where

$$\mathbf{F}' = \begin{bmatrix} .344 & .353 & .380 & .397 & .371 & .362 & .433 \\ .473 & .353 & -.006 & -.246 & .113 & .233 & -.724 \end{bmatrix}.$$

Note, once again, that $\boldsymbol{\gamma}_1 \in \mathcal{N}\,(\mathbf{C})$, and, hence, the requirements for Theorem 2.5 are met.

For the first population, we now have reduced two-dimensional parameters

$$\Pi_1 : \boldsymbol{\mu}_{1R} = \begin{bmatrix} 10.41 \\ 3.63 \end{bmatrix}, \boldsymbol{\Sigma}_{1R} = \begin{bmatrix} 3.69 & -.037 \\ -.037 & 2.99 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{1R} = \begin{bmatrix} 1.86 \\ -.028 \end{bmatrix}.$$

Also, for the second population, we have reduced two-dimensional parameters

$$\Pi_2 : \boldsymbol{\mu}_{2R} = \begin{bmatrix} 15.53 \\ 7.53 \end{bmatrix}, \boldsymbol{\Sigma}_{2R} = \begin{bmatrix} 3.06 & -.004 \\ -.004 & 2.99 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{2R} = \begin{bmatrix} 1.60 \\ -.014 \end{bmatrix}.$$

Finally, for the third population, we have reduced two-dimensional parameters

$$\Pi_3 : \boldsymbol{\mu}_{3R} = \begin{bmatrix} 18.33 \\ 3.21 \end{bmatrix}, \boldsymbol{\Sigma}_{3R} = \begin{bmatrix} 9.97 & -.37 \\ -.37 & 3.01 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{3R} = \begin{bmatrix} 2.92 \\ -.083 \end{bmatrix}.$$

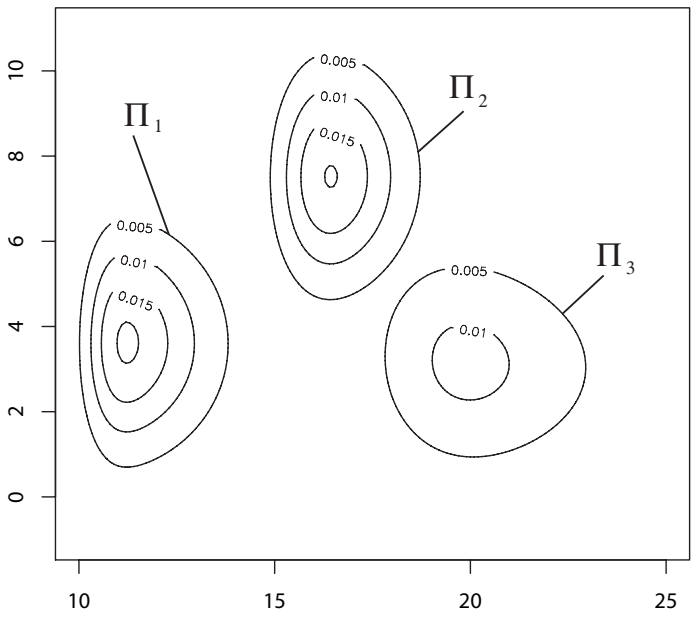

Figure 2.4: The optimal two-dimensional representation for the three seven-dimensional skew-normal populations in Example 2.3.

Immediately above, we portray a two-dimensional representation of the original seven-dimensional MSN populations. A one-dimensional representation of these three populations is obtainable if we use the SVD described in Theorem 2.6; however, we lose some discriminatory information by applying this procedure. For the first population, we obtain the reduced one-dimensional parameters

$$\Pi_1 : \mu_{1R} = 10.41, \sigma_{1R} = 3.69, \text{ and } \gamma_{1R} = 1.86.$$

For the second population, we obtain the reduced one-dimensional parameters

$$\Pi_2 : \mu_{2R} = 15.53, \sigma_{2R} = 3.06, \text{ and } \gamma_{2R} = 1.60.$$

The reduced one-dimensional parameters for the third population are

$$\Pi_3 : \mu_{3R} = 18.33, \sigma_{3R} = 9.97, \text{ and } \gamma_{3R} = 2.92.$$

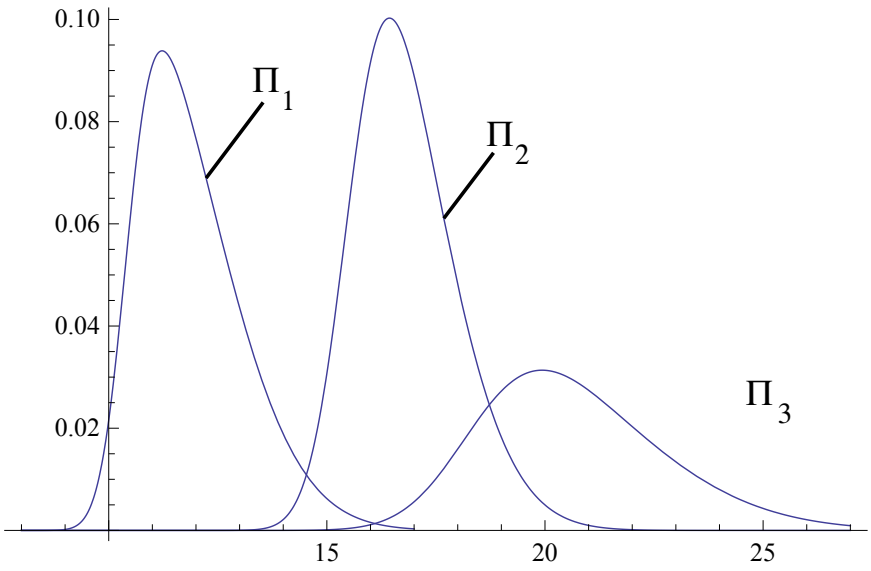

Figure 2.5: A one-dimensional approximate representation for the three skew-normal populations in Example 2.3.

2.6 Concluding Remarks

We have presented a simple and flexible algorithm for low-dimensional representation of data from several *MSN* populations under different parametric configurations with possibly unequal dispersion and skew parameters. Necessary and sufficient conditions have been given for attaining the smallest dimensional subspace $q \ll p$ that preserves the original Bayes classification assignments. Also, we have given a constructive proof for obtaining a low-dimensional representation space for multiple *MSN* densities when certain conditions are satisfied.

In addition, we remark that the restriction $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$ is not as limiting as one might initially believe. If $\boldsymbol{\gamma}_1 \notin \mathcal{N}(\mathbf{C})$ or Corollary 2.2 and Corollary 2.3 do not hold, we can perform a Box-Cox transformation on the m populations so that at least one population is represented by an approximate multivariate normal population. We can then apply Corollary 2.2 or 2.3, depending on the parameter configuration.

We note several advantages of our newly proposed low-dimensional representation or *LDR* method given in Theorems 2.2-2.5. First, the method is not restricted to

a one-dimensional representation regardless of the number of populations. Second, the method allows for equal and unequal covariance structures. Third, the original feature dimension p does not significantly impact the computational complexity. Finally, the skewness parameters allow for density functions that are shaped much differently from multivariate normal densities.

CHAPTER THREE

Linear Dimension Reduction for Multiple Multivariate Singular Skew-Normal Densities

3.1 Introduction

The theorems from Chapter 2 hold true provided we have a multivariate skew-normal distribution with nonsingular dispersion parameter $\mathbf{\Sigma}_i$, $i = 1, 2, ..., m$. However, the dispersion parameter may be singular. In this chapter, we give necessary and sufficient conditions for which a low-dimensional linear transformation of the original data will preserve the expected Bayes probability of misclassification in the original measurement space when the populations have a multivariate skew-normal distribution with singular dispersion parameters $\mathbf{\Sigma}_i$, $i = 1, 2, ..., m$. In addition, we give a method for construction of this linear dimension-compression matrix.

It is a well-known fact that the density of a normally distributed vector with singular covariance matrix does not exist with respect to the Lebesgue measure on $\mathbb{R}^p$. However, Khatri (1968) has shown that a multivariate normal density function does exist on a subspace of $\mathbb{R}^p$. Van Perlo-ten Kleij (2004) has defined the multivariate singular normal distribution as follows:

Definition 3.1. If $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ with $rank(\mathbf{\Sigma}) = k < p$, then $\mathbf{x}$ has the probability density function

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} \left[\prod_{i=1}^{r} \lambda_i(\mathbf{A})\right]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \mathbf{\Sigma}^{+}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{3.1}$$

for $\mathbf{x} \in \mathbf{V} = \boldsymbol{\mu} + K(\mathbf{\Sigma})^{\perp}$ with respect to the Lebesgue measure $\lambda_{\mathcal{V}}$ on an affine subspace $\mathbf{V}$ of $\mathbb{R}^p$ of dimension k. As in Chapter 2, the notation $\mathbf{\Sigma}^{+}$ represents the Moore-Penrose pseudoinverse of $\mathbf{\Sigma}$.

Let $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$ with $rank(\boldsymbol{\Sigma}) = k < p$. Then, the null space of $\boldsymbol{\Sigma}$ has $dim[\mathcal{N}(\boldsymbol{\Sigma})] = p - k$, and the random vector $\mathbf{x}$ will have outcomes in an affine subspace of $\mathbb{R}^m$. We now present a derivation to prove the existence of the multivariate skew-normal ($MSSN$) density. Our derivation is similar to one used by van Perlo-ten Kleij (2004) for the singular multivariate normal density.

Proposition 3.1. Let $\mathbf{x} \sim SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$ with $rank(\boldsymbol{\Sigma}) = k < p$. Then, $\mathbf{x} \in \mathcal{V}$ with probability one, where $\mathcal{V} \equiv \boldsymbol{\mu} + \mathcal{N}(\boldsymbol{\Sigma})^{\perp}$.

Clearly, $\mathcal{V}$ is an affine subspace of $\mathbb{R}^p$ with dimension k. Now, let $\boldsymbol{\Sigma} = \mathbf{C}\boldsymbol{\Lambda}\mathbf{C}'$ be a spectral decomposition of $\boldsymbol{\Sigma}$ with

$$\boldsymbol{\Lambda} = diag(\lambda_1, ..., \lambda_k, 0, ..., 0) = \begin{bmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where $\boldsymbol{\Lambda}_1 = diag(\lambda_1, ..., \lambda_k)$, $\lambda_1 \geq ... \geq \lambda_k > 0$, and $\mathbf{C}$ is an orthogonal $p \times p$ matrix. Moreover, let

$$\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2],$$

where $\mathbf{C}_1 = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k]$ and $\mathbf{C}_2 = [\mathbf{c}_{k+1}, \mathbf{c}_{k+2}, ..., \mathbf{c}_p]$. Hence, $\{\mathbf{c}_1, ..., \mathbf{c}_m\}$ is an orthonormal basis for $\mathbb{R}^m$ such that $\{\mathbf{c}_1, ..., \mathbf{c}_k\}$ is an orthonormal basis for $\mathcal{N}(\boldsymbol{\Sigma})^{\perp}$ and $\{\mathbf{c}_{k+1}, ..., \mathbf{c}_p\}$ is an orthonormal basis for $\mathcal{N}(\boldsymbol{\Sigma})$. Furthermore, $\mathcal{C}(\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\Sigma})^{\perp}$, where $\mathcal{C}(\boldsymbol{\Sigma})$ denotes the column space of $\boldsymbol{\Sigma}$. Then,

$$\begin{aligned} \mathbf{x} \in \mathcal{C}^{\perp}(\boldsymbol{\Sigma}) &\Leftrightarrow \mathbf{x}'\mathbf{y} = 0 \text{ for all } \mathbf{y} \in \mathcal{C}(\boldsymbol{\Sigma}) \\ &\Leftrightarrow \mathbf{x}'\boldsymbol{\Sigma}\mathbf{z} = 0 \text{ for all } \mathbf{z} \in \mathbb{R}^k \\ &\Leftrightarrow \boldsymbol{\Sigma}\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \mathbf{x} \in \mathcal{N}(\boldsymbol{\Sigma}). \end{aligned}$$

Proof: We observe that

$$\begin{aligned} P(\mathbf{x} \in \mathcal{V}) &= P\left((\mathbf{x}-\boldsymbol{\mu}) \in \mathcal{N}(\boldsymbol{\Sigma})^{\perp}\right) \\ &= P(\mathbf{y}'(\mathbf{x}-\boldsymbol{\mu}) = 0 \text{ for all } \mathbf{y} \in \mathcal{N}(\boldsymbol{\Sigma})) \\ &= P\left(\mathbf{c}_j'(\mathbf{x}-\boldsymbol{\mu}) = 0 \text{ for } j = k+1, ..., p\right). \end{aligned}$$

Then, for $j = k+1, ..., p$,

$$\mathbf{c}_j'(\mathbf{X}-\boldsymbol{\mu}) \sim SSN\left(0, \mathbf{c}_j'\boldsymbol{\Sigma}\mathbf{c}_j, \delta_0, \mathbf{c}_j'\boldsymbol{\gamma}\right) = \delta_{\{0\}},$$

where $\delta_{\{0\}}$ denotes a degenerate random variable. Therefore,

$$\begin{aligned} P(\mathbf{x} \notin \mathcal{V}) &= P\left(\mathbf{c}_j'(\mathbf{x}-\boldsymbol{\mu}) \neq 0 \text{ for some } j = k+1, ..., p\right) \\ &\leq \sum_{j=k+1}^{m} P\left(\mathbf{c}_j'(\mathbf{x}-\boldsymbol{\mu}) \neq 0\right) \\ &= 0, \end{aligned}$$

which implies $P(\mathbf{x} \in \mathcal{V}) = 1$.

Now we derive the probability density function of $\mathbf{x}$ with respect to the Lebesgue measure $\lambda_{\mathcal{V}}$ on $\mathcal{V}$. Consider the affine transformation $T : \mathbb{R}^k \to \mathcal{V}$ defined by

$$\mathbf{x} = \mathbf{C}_1\mathbf{y} + \boldsymbol{\mu}.$$

Note that $\mathbf{y} = T^{-1}(\mathbf{x}) = \mathbf{C}_1'(\mathbf{x}-\boldsymbol{\mu})$. Let λ_k denote the Lebesgue measure on $\mathbb{R}^k$. The following properties hold for T:

(i) T is onto and one-to-one;

(ii) T is bicontinuous;

(iii) $\lambda_{\mathcal{V}} = \lambda_k T^{-1}$.

Property (iii) holds for all $\boldsymbol{\mu} \in \mathcal{V}$ and for all orthonormal bases $\mathbf{c}_1, ..., \mathbf{c}_k$ for the subspace $\mathcal{V} - \boldsymbol{\mu}$. For the following theorem and for the remainder of the chapter, we use

$$det_r\mathbf{A} \equiv \prod_{i=1}^{r} \lambda_i(\mathbf{A}) \tag{3.2}$$

to denote the product of the r nonzero eigenvalues of $\mathbf{A} \in \mathbb{R}^S_{n\times n}$ such that $rank(\mathbf{A}) = r$.

Theorem 3.1. If $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$ with $rank(\boldsymbol{\Sigma}) = k < p$, then $\mathbf{x}$ has the probability density function

$$f(\mathbf{x}) = \frac{1}{\Phi(\delta_0)} (2\pi)^{-\frac{k}{2}} [det_k \boldsymbol{\Sigma}]^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^+ (\mathbf{x}-\boldsymbol{\mu})\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^+(\mathbf{x}-\boldsymbol{\mu})}{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}}\right) \tag{3.3}$$

for $\mathbf{x} \in \boldsymbol{\mu} + \mathcal{N}(\boldsymbol{\Sigma})^\perp$ with respect to the Lebesgue measure $\lambda_{\mathcal{V}}$, where $det_k \boldsymbol{\Sigma}$ is defined in (3.2).

Proof: Let $\mathbf{y} = \mathbf{C}_1'(\mathbf{x}-\boldsymbol{\mu})$, where $\mathbf{y} \sim SN_m(\mathbf{0}, \boldsymbol{\Lambda}_1, \delta_0, \boldsymbol{\tau})$ and $\boldsymbol{\tau} = \mathbf{C}_1'\boldsymbol{\gamma}$. From the characteristic function of $\mathbf{y}$, we know that $y_1, ..., y_k$ are independently distributed with $y_i \sim SN_m(\mathbf{0}, \lambda_i, \delta_0, \boldsymbol{\tau})$ for $i = 1, ..., k$. This fact implies that $\mathbf{y}$ has density function

$$\begin{aligned} h(\mathbf{y}) &= \prod_{i=1}^{k} \frac{1}{\Phi(\delta_{0i})} (2\pi)^{-\frac{1}{2}} \lambda_i^{-\frac{1}{2}} \exp\left(-\frac{1}{2} y_i^2/\lambda_i\right) \Phi\left(\frac{\delta_{0i} + (\tau_i^2)\, y_i/\lambda_i}{\sqrt{1-\tau_i^2/\lambda_i}}\right) \\ &= \frac{1}{\Phi(\delta_0)} (2\pi)^{-\frac{k}{2}} [det_k \boldsymbol{\Lambda}_1]^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{y}'\boldsymbol{\Lambda}_1^{-1}\mathbf{y}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\tau}'\boldsymbol{\Lambda}_1^{-1}\mathbf{y}}{\sqrt{1-\boldsymbol{\tau}'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\tau}}}\right) \end{aligned}$$

with respect to the Lebesgue measure λ_k on $\mathbb{R}^k$, where $\tau_i^2/\lambda_i < 1$ and $\boldsymbol{\tau}'\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\tau} < 1$. We have $\mathbf{y} = T^{-1}(\mathbf{x})$. Let B be any measurable set in $\mathcal{V}$. Then,

$$\begin{aligned} P(\mathbf{x} \in B) &= P\left(T^{-1}(\mathbf{x}) \in T^{-1}(B)\right) \\ &= P\left(\mathbf{y} \in T^{-1}(B)\right) \\ &= \int_{T^{-1}(B)} h d\lambda_k \\ &= \int_{T^{-1}(B)} h d\lambda_{\mathcal{V}} T \\ &= \int_B h\left(T^{-1}(\mathbf{x})\right) d\lambda_{\mathcal{V}}(\mathbf{x}). \end{aligned}$$

Also, note that

$$\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{+}\boldsymbol{\gamma} = \boldsymbol{\gamma}' \begin{bmatrix} \mathbf{C}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \boldsymbol{\gamma}$$
$$= \boldsymbol{\gamma}'\mathbf{C}_1'\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1\boldsymbol{\gamma}.$$

Then, a representation of the $MSSN$ density function is

$$\begin{aligned} f(\mathbf{x}) &= h\left(T^{-1}(\mathbf{x})\right) \\ &= \frac{1}{\Phi(\delta_0)}(2\pi)^{-\frac{k}{2}}\left[det_k\boldsymbol{\Lambda}_1\right]^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\mathbf{C}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1'(\mathbf{x}-\boldsymbol{\mu})\right)\times \\ &\quad \Phi\left(\frac{\delta_0+\boldsymbol{\gamma}'\mathbf{C}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1'}{\sqrt{1-\boldsymbol{\gamma}'\mathbf{C}_1\boldsymbol{\Lambda}_1^{-1}\mathbf{C}_1'\boldsymbol{\gamma}}}\right) \\ &= \frac{1}{\Phi(\delta_0)}(2\pi)^{-\frac{k}{2}}\left[det_k\boldsymbol{\Sigma}\right]^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{+}(\mathbf{x}-\boldsymbol{\mu})\right)\times \\ &\quad \Phi\left(\frac{\delta_0+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{+}(\mathbf{x}-\boldsymbol{\mu})}{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{+}\boldsymbol{\gamma}}}\right), \end{aligned} \tag{3.4}$$

which is a density function of $\mathbf{x}$ on $\mathcal{V}$ with respect to $\lambda_{\mathcal{V}}$. If we take $k = p$ in (3.4), the probability density function (3.4) becomes the regular $MSSN$ density function (2.1) with respect to the Lebesgue measure on $\mathbb{R}^p$.

Definition 3.2. If $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_o, \boldsymbol{\gamma})$ with $rank(\boldsymbol{\Sigma}) = k < p$, then $\mathbf{x}$ has the probability density function

$$p(\mathbf{x}|\Pi) = \frac{1}{\Phi(\delta_0)}\varphi_p^S(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{+}(\mathbf{x}-\boldsymbol{\mu})}{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^{+}\boldsymbol{\gamma}}}\right), \tag{3.5}$$

where $\mathbf{x} \in \mathbb{R}_{p\times 1}$, $\boldsymbol{\mu} \in \mathbb{R}_{p\times 1}$, $\boldsymbol{\Sigma} \in \mathbb{R}_{p\times p}^{\geq}$, $\boldsymbol{\gamma} \in \mathbb{R}_{p\times 1}$, $\delta_0 \in \mathbb{R}$, and $\varphi_p^S(\mathbf{x})$ and $\Phi(\mathbf{x})$ denote the p-dimensional singular normal density function defined in (3.1) and the univariate standard normal distribution function, respectively.

Here, we derive the singular skew-normal moment-generating function.

Theorem 3.2. Let $\mathbf{v} \sim SN_p(\mathbf{0}, \mathbf{I}_p, \delta_0, \boldsymbol{\gamma})$ and $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, where $rank(\boldsymbol{\Sigma}) = k < p$ with $\mathbf{v} = \left[\boldsymbol{\Sigma}^{\frac{1}{2}}\right]^{+}(\mathbf{x}-\boldsymbol{\mu})$. Hence,

$$M_{\mathbf{v}}(\mathbf{t}) = \frac{1}{\Phi(\delta_0)}\exp\left\{\frac{\mathbf{t}'\mathbf{t}}{2}\right\}\Phi\left(\frac{\lambda_0+\boldsymbol{\lambda}_1'\mathbf{t}}{\sqrt{1+\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}\right),$$

where

$$\delta_0 = \frac{\lambda_0}{\sqrt{1+\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}, \boldsymbol{\lambda}_1 = \frac{\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\right]^+\boldsymbol{\gamma}}{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}}, \text{ and } 1+\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1 = 1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}.$$

Also,

$$\frac{\boldsymbol{\lambda}_1'\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{\sqrt{1+\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}} = \frac{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}}{1}\frac{\boldsymbol{\gamma}'\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\right]^+\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}} = \boldsymbol{\gamma}'\mathbf{t},$$

where $\boldsymbol{\gamma} \in \mathcal{C}(\boldsymbol{\Sigma})$. Then,

$$\begin{aligned}
M_{\mathbf{x}}(\mathbf{t}) &= M_{\boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\mathbf{v}}(\mathbf{t}) \\
&= \exp\{\mathbf{t}'\boldsymbol{\mu}\} M_{\mathbf{v}}\left(\boldsymbol{\Sigma}^{1/2}\mathbf{t}\right) \\
&= \exp\{\mathbf{t}'\boldsymbol{\mu}\}\frac{1}{\Phi(\delta_0)}\exp\left\{\frac{\mathbf{t}'\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{2}\right\}\Phi\left(\frac{\lambda_0+\boldsymbol{\lambda}_1'\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}}{\sqrt{1+\boldsymbol{\lambda}_1'\boldsymbol{\lambda}_1}}\right) \\
&= \frac{1}{\Phi(\delta_0)}\exp\left\{\mathbf{t}'\boldsymbol{\mu}+\frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\right\}\Phi\left(\delta_0+\frac{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}}{\sqrt{1-\boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}}\boldsymbol{\gamma}'\left[\boldsymbol{\Sigma}^{\frac{1}{2}}\right]^+\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{t}\right) \\
&= \exp\left\{\mathbf{t}'\boldsymbol{\mu}+\frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}\right\}\frac{\Phi(\delta_0+\boldsymbol{\gamma}'\mathbf{t})}{\Phi(\delta_0)},
\end{aligned}$$

where $\boldsymbol{\gamma} \in \mathcal{C}(\boldsymbol{\Sigma})$.

We now give two fundamental properties for the *MSSN* random vector. Proposition 3.2 provides the linear combination of a $MSSN$ density function. Next, Proposition 3.3 gives the marginal density functions for the $MSSN$ density function. Also, Proposition 3.4 gives the conditional density function of the $MSSN$ density function.

Proposition 3.2. Let $\mathbf{b} \in \mathbb{R}_{m\times 1}$ and $\mathbf{C} \in \mathbb{R}_{m\times p}$ with $rank(\mathbf{C}) = m$, where $m \leq p$. If $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, where $rank(\boldsymbol{\Sigma}) = k < p$, then

$$\mathbf{C}\mathbf{x}+\mathbf{b} \sim SSN_m(\mathbf{C}\boldsymbol{\mu}+\mathbf{b}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}', \delta_0, \mathbf{C}\boldsymbol{\gamma}).$$

Proof: For $\mathbf{t} \in \mathbb{R}^p$, the moment-generating function of $\mathbf{C}\mathbf{x}$ is given by

$$\begin{aligned}
M_{\mathbf{b}+\mathbf{C}\mathbf{x}}(\mathbf{t}) &= \exp^{\mathbf{t}'\mathbf{b}} M_{\mathbf{x}}(\mathbf{C}'\mathbf{t}) \\
&= \exp\left\{\mathbf{t}'\mathbf{C}\boldsymbol{\mu}+\mathbf{t}'\mathbf{b}+\frac{\mathbf{t}'\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\mathbf{t}}{2}\right\}\frac{\Phi(\delta_0+\boldsymbol{\gamma}'\mathbf{C}'\mathbf{t})}{\Phi(\delta_0)}. \qquad (3.6)
\end{aligned}$$

Because (3.6) is the moment-generating function of a $MSSN$ distribution with $rank\,(\boldsymbol{\Sigma}) = k < p$, we conclude that $\mathbf{Ax} \sim SSN_m\,(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}', \delta_0, \mathbf{A}\boldsymbol{\gamma})$.

Proposition 3.3. Let $\mathbf{x} \sim SSN_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$. If we partition $\mathbf{x} = [\mathbf{x}_1', \mathbf{x}_2']'$ into two subvectors of dimensions m and $p-m$, respectively, and correspondingly partition

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \text{ and } \boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix}, \tag{3.7}$$

then

(i) $\mathbf{x}_1 \sim SSN_m\,(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \delta_0, \boldsymbol{\gamma}_1)$;

(ii) $\mathbf{x}_2 \sim SSN_{p-m}\,(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}, \delta_0, \boldsymbol{\gamma}_2)$.

Proof: We will use the moment-generating function. Taking $\mathbf{t}_2 = \mathbf{0}$, we get

$$M_{\mathbf{x}} \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{0} \end{pmatrix} = \exp\left\{\mathbf{t}_1'\boldsymbol{\mu}_1 + \frac{\mathbf{t}_1'\boldsymbol{\Sigma}_{11}\mathbf{t}_1}{2}\right\} \frac{\Phi\,(\delta_0 + \boldsymbol{\gamma}_1'\mathbf{t}_1)}{\Phi\,(\delta_0)}$$

in order to obtain (i). In order to show (ii), we set $\mathbf{t}_1 = \mathbf{0}$ and see that

$$M_{\mathbf{x}} \begin{pmatrix} \mathbf{0} \\ \mathbf{t}_2 \end{pmatrix} = \exp\left\{\mathbf{t}_2'\boldsymbol{\mu}_2 + \frac{\mathbf{t}_2'\boldsymbol{\Sigma}_{22}\mathbf{t}_2}{2}\right\} \frac{\Phi\,(\delta_0 + \boldsymbol{\gamma}_2'\mathbf{t}_2)}{\Phi\,(\delta_0)}.$$

Proposition 3.4. Let $\mathbf{x} \sim SSN_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$. If we partition $\mathbf{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\gamma}$ as in (3.7), then the conditional density function of $\mathbf{x}_1|\mathbf{x}_2$ is

$$g\,(\mathbf{x}_1|\mathbf{x}_2) = \frac{1}{\Phi\,(\beta_0)} \varphi_m^S\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right) \Phi\left(l_0 + \mathbf{l}_1'\left(\boldsymbol{\Sigma}_{1|2}^{\frac{1}{2}}\right)^{+}\left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right)\right),$$

where

$$\boldsymbol{\mu}_{1|2} \equiv \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{+}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right), \tag{3.8}$$

$$\boldsymbol{\Sigma}_{1|2} \equiv \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{+}\boldsymbol{\Sigma}_{21}, \tag{3.9}$$

$$l_0 \equiv \frac{\delta_0 + \boldsymbol{\gamma}_2'\boldsymbol{\Sigma}_{22}^{+}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^{+}\boldsymbol{\gamma}}}, \tag{3.10}$$

$$\beta_0 \equiv \frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} \boldsymbol{\gamma}_2}}, \tag{3.11}$$

and

$$\mathbf{l}_1 \equiv \frac{\left[\left(\boldsymbol{\Sigma}_{1|2}\right)^{\frac{1}{2}}\right]^{+} \left(\boldsymbol{\gamma}_1 - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{+} \boldsymbol{\gamma}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} \boldsymbol{\gamma}}}. \tag{3.12}$$

Proof: From Definition 3.2, the joint density function of $\mathbf{x}$ is

$$f\left(\mathbf{x}\right) = \frac{1}{\Phi\left(\delta_0\right)} \varphi_p^S\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} \left(\mathbf{x} - \boldsymbol{\mu}\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} \boldsymbol{\gamma}}}\right).$$

Also, Proposition 3.3 states that the marginal density function of $\mathbf{x}_2$ is

$$f\left(\mathbf{x}_2\right) = \frac{1}{\Phi\left(\delta_0\right)} \varphi_{p-m}^S\left(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} \boldsymbol{\gamma}_2}}\right).$$

By the definition of the conditional density function of $\mathbf{x}_1$ given $\mathbf{x}_2$, we have that

$$\begin{aligned}
f\left(\mathbf{x}_1 | \mathbf{x}_2\right) &= \frac{\frac{1}{\Phi(\delta_0)} \varphi_p^S\left(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} (\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} \boldsymbol{\gamma}}}\right)}{\frac{1}{\Phi(\delta_0)} \varphi_{p-m}^S\left(\mathbf{x}_2; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} (\mathbf{x}_2 - \boldsymbol{\mu}_2)}{\sqrt{1 - \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} \boldsymbol{\gamma}_2}}\right)} \\
&= \frac{1}{\Phi\left(\beta_0\right)} \varphi_m^S\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right) \times \\
&\quad \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_{22}^{+} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} \boldsymbol{\gamma}}} + \frac{\left(\boldsymbol{\gamma}_1 - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{+} \boldsymbol{\gamma}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}^{+} \boldsymbol{\gamma}}} \boldsymbol{\Sigma}_{1|2}^{+} \left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right)\right) \\
&= \frac{1}{\Phi\left(\beta_0\right)} \varphi_m^S\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right) \Phi\left(l_0 + \mathbf{l}_1' \left(\boldsymbol{\Sigma}_{1|2}^{\frac{1}{2}}\right)^{+} \left(\mathbf{x}_1 - \boldsymbol{\mu}_{1|2}\right)\right),
\end{aligned}$$

where $\boldsymbol{\mu}_{1|2}$, $\boldsymbol{\Sigma}_{1|2}$, l_0, β_0, and $\mathbf{l}_1$ are given in (3.8) through (3.12), respectively.

Because (3.5) is positive and the logarithm function of (3.5) is monotonic increasing, the Bayes decision rule for classifying $\mathbf{x}$ for the two-population $MSSN$ case can be found explicitly. The Bayes decision rule is:

Assign $\mathbf{x}$ to Π_1 if

$$\begin{aligned}
&\frac{1}{\Phi\left(\delta_0\right)} \varphi_p^S\left(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_1' \boldsymbol{\Sigma}_1^{+} \left(\mathbf{x} - \boldsymbol{\mu}_1\right)}{\sqrt{1 - \boldsymbol{\gamma}_1' \boldsymbol{\Sigma}_1^{+} \boldsymbol{\gamma}_1}}\right) > \\
&\qquad \frac{1}{\Phi\left(\delta_0\right)} \varphi_p^S\left(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_2^{+} \left(\mathbf{x} - \boldsymbol{\mu}_2\right)}{\sqrt{1 - \boldsymbol{\gamma}_2' \boldsymbol{\Sigma}_2^{+} \boldsymbol{\gamma}_2}}\right)
\end{aligned}$$

and to Π_2, otherwise. For the case of $m > 2$ classes, the Bayes classification rule is to classify the unlabeled observation vector $\mathbf{x}$ into class Π_k corresponding to the minimum distance function $d_k^{SN} = \min\{p(\mathbf{x}|\Pi_i),\ i = 1, 2, ..., m\}$.

3.2 Preliminary Results

The proof of our new linear dimension reduction theorem for $MSSN$ densities requires the following notation and lemmas. Consider $\mathbf{M} \in \mathbb{R}_{p\times(m-1)(p+1)}$, where

$$\mathbf{M} \equiv \left[\mathbf{E}_2^+(\mathbf{d}_2 - \mathbf{d}_1)|...|\mathbf{E}_m^+(\mathbf{d}_m - \mathbf{d}_1)|\mathbf{E}_2 - \mathbf{E}_1|...|\mathbf{E}_m - \mathbf{E}_1|\mathbf{a}_2 - \mathbf{a}_1|...|\mathbf{a}_m - \mathbf{a}_1\right],$$

where $\mathbf{a}_i, \mathbf{d}_i \in \mathbb{R}_{p\times 1}$, $\mathbf{E}_i \in \mathbb{R}_{p\times p}^S$ with $rank(\mathbf{E}_i) = k < p$ for $i = 1, 2, ..., m$ and $\mathbf{E}_1 \neq \mathbf{E}_k$ for at least one value of k, where $2 \leq k \leq m$. Also, let $rank(\mathbf{M}) = 1 \leq q < p$, and let $\mathbf{M} = \mathbf{FG}$, where $\mathbf{F} \in \mathbb{R}_{p\times q}$ and $\mathbf{G} \in \mathbb{R}_{q\times(m-1)(p+1)}$ with $rank(\mathbf{F}) = rank(\mathbf{G}) = q$. Then, the Moore-Penrose pseudoinverse of $\mathbf{M}$ is $\mathbf{M}^+ = \mathbf{G}^+\mathbf{F}^+$, and we have that $\mathbf{MM}^+ = \mathbf{FGG}^+\mathbf{F}^+ = \mathbf{FF}^+$ and $\mathbf{MM}^+\mathbf{M} = \mathbf{FF}^+\mathbf{M} = \mathbf{M}$. This property implies that for $i = 1, 2, ..., m$,

(1) $\mathbf{FF}^+(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{a}_i - \mathbf{a}_1$,

(2) $\mathbf{FF}^+(\mathbf{E}_i - \mathbf{E}_1) = \mathbf{E}_i - \mathbf{E}_1$, and

(3) $\mathbf{FF}^+\left[\mathbf{E}_i^+(\mathbf{d}_i - \mathbf{d}_1)\right] = \mathbf{E}_i^+(\mathbf{d}_i - \mathbf{d}_1)$.

We now give three lemmas used in the proof of our main result. The reader can find the proofs of Lemmas 3.1 and 3.2 in Onsupreth and Young (2005).

Lemma 3.1. Let $\mathbf{F} \in \mathbb{R}_{p\times q}$ and $\mathbf{C} = \mathbf{R}[\mathbf{I} - \mathbf{FF}^+]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{C}) = p - q$ and $\mathbf{E}_i \in \mathbb{R}_{p\times p}^S$ with $rank(\mathbf{E}_i) < p$ for $i = 1, 2, ..., m$, such that properties (1) - (3) hold. Then,

(a) $\mathbf{FF}^+(\mathbf{E}_i - \mathbf{E}_1) = (\mathbf{E}_i - \mathbf{E}_1)\mathbf{FF}^+$,

(b) $\mathbf{FF}^+\mathbf{E}_i = \mathbf{E}_i\mathbf{FF}^+$,

(c) $(\mathbf{I} - \mathbf{FF}^+)\mathbf{E}_i = \mathbf{E}_1(\mathbf{I} - \mathbf{FF}^+)$, and

(d) $\mathbf{CE}_i\mathbf{C}' = \mathbf{CE}_1\mathbf{C}'$.

The following lemma provides the pseudoinverse for the quantity $\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}$.

Lemma 3.2. Let $\mathbf{F} \in \mathbb{R}_{p\times q}$ and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\left(\mathbf{E}_i\right) < p$ for $i = 1, 2, ..., m$, such that properties (1) - (3) hold. Then, $\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+} = \mathbf{F}'\mathbf{E}_i^{+}\mathbf{F}$.

Lemma 3.3. Let $\mathbf{F} \in \mathbb{R}_{p\times q}$, $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\left(\mathbf{E}_i\right) < p$, and $\mathbf{C} = \mathbf{R}[\mathbf{I} - \mathbf{F}\mathbf{F}^{+}]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{C}) = p - q$ for $i = 1, 2, ..., m$, such that properties (1) - (3) hold. Then, $\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime} = \mathbf{0}$.

Proof: By Property (2), we have that

$$\begin{aligned}\mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime} &= \mathbf{R}\left(\mathbf{I} - \mathbf{F}\mathbf{F}^{+}\right)\mathbf{E}_i\mathbf{F}^{+\prime} \\ &= \mathbf{R}\left(\mathbf{E}_i - \mathbf{F}\mathbf{F}^{+}\mathbf{E}_i\right)\mathbf{F}^{+\prime} \\ &= \mathbf{0}.\end{aligned}$$

Lemma 3.4. Let $\mathbf{a}_i, \mathbf{d}_i \in \mathbb{R}_{p\times 1}$, $\mathbf{F} \in \mathbb{R}_{p\times q}$, and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\left(\mathbf{E}_i\right) < p$ for $i = 1, 2, ..., m$ such that properties (1) - (3) hold, and let $\mathbf{C} = \mathbf{R}\left[\mathbf{I} - \mathbf{F}\mathbf{F}^{+}\right]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank\left(\mathbf{C}\right) = p - q$. Then, for $i = 1, 2, ..., m$,

(a) $\mathbf{C}\mathbf{F} = \mathbf{0}$,

(b) $\mathbf{C}\mathbf{d}_i = \mathbf{C}\mathbf{d}_1$,

(c) $\mathbf{C}\mathbf{a}_i = \mathbf{C}\mathbf{a}_1$,

(d) $\mathbf{C}\mathbf{d}_i + \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+}\left(\mathbf{y} - \mathbf{F}^{+}\mathbf{d}_i\right) = \mathbf{C}\mathbf{d}_1$, where $\mathbf{y} \in \mathcal{C}\left(\mathbf{E}_i\right)$, $i = 2, 3, ..., m$,

(e) $\mathbf{C}\mathbf{E}_i\mathbf{C}' - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}' = \mathbf{C}\mathbf{E}_1\mathbf{C}'$,

(f) $\mathbf{a}_i'\mathbf{C}'\left[\mathbf{C}\mathbf{E}_i\mathbf{C}' - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right]^{+}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{d}_1\right)$
$= \left(\mathbf{C}\mathbf{a}_1\right)'\left(\mathbf{C}\mathbf{E}_1\mathbf{C}'\right)^{+}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{d}_1\right)$,

(g) $\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\left[\mathbf{C}\mathbf{E}_i\mathbf{C}' - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+}\mathbf{F}^{+}\mathbf{E}_i\mathbf{C}'\right]^{+}\mathbf{C}\mathbf{a}_i = \mathbf{0}$,

(h) $\mathbf{C}\mathbf{a}_i - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+}\mathbf{F}^{+}\mathbf{a}_i = \mathbf{C}\mathbf{a}_1$.

Proof of (a). The proof of (a) is obvious and is not given.

Proof of (b). Let $(\mathbf{d}_i - \mathbf{d}_1) \in \mathcal{C}(\mathbf{E}_i)$, $i = 2, 3, ..., m$. Then,

$$\begin{aligned}\mathbf{F}\mathbf{F}^+\left[\mathbf{E}_i^+(\mathbf{d}_i - \mathbf{d}_1)\right] = \mathbf{E}_i^+(\mathbf{d}_i - \mathbf{d}_1) &\Rightarrow \mathbf{E}_i\left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)\mathbf{E}_i^+(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{0} \\ &\Rightarrow \left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)\mathbf{E}_i\mathbf{E}_i^+(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{0} \\ &\Rightarrow \mathbf{R}\left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)(\mathbf{d}_i - \mathbf{d}_1) = \mathbf{0} \\ &\Rightarrow \mathbf{C}\mathbf{d}_i = \mathbf{C}\mathbf{d}_1.\end{aligned}$$

Proof of (c). Because $(\mathbf{a}_i - \mathbf{a}_1) \in \mathcal{C}(\mathbf{F})$, we have that

$$\begin{aligned}\mathbf{F}\mathbf{F}^+(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{a}_i - \mathbf{a}_1 &\Rightarrow \left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{0} \\ &\Rightarrow \mathbf{R}\left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)(\mathbf{a}_i - \mathbf{a}_1) = \mathbf{0} \\ &\Rightarrow \mathbf{C}\mathbf{a}_i = \mathbf{C}\mathbf{a}_1.\end{aligned}$$

Proof of (d). From Lemma 3.3, we have

$$\left(\mathbf{C}\mathbf{E}_i\mathbf{F}^{+'}\right)\left(\mathbf{F}^+\mathbf{E}_i\mathbf{F}^{+'}\right)'\mathbf{F}^+\mathbf{E}_i\mathbf{C}' = \mathbf{0},$$

and from Lemma 3.1.d, we see that $\mathbf{C}\mathbf{E}_i\mathbf{C}' = \mathbf{C}\mathbf{E}_1\mathbf{C}'$.

Proof of (e). Part (e) follows as a direct result of Lemmas 3.1.d and 3.3.

Proof of (f). From Lemmas 3.3 and 3.4.b, we have

$$\begin{aligned}&(\mathbf{C}\mathbf{a}_i)'\left[\mathbf{C}\mathbf{E}_i\mathbf{C}' - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+'}\left(\mathbf{F}^+\mathbf{E}_i\mathbf{F}^{+'}\right)^+\mathbf{F}^+\mathbf{E}_i\mathbf{C}'\right]^+(\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{a}_1) \\ &\quad = (\mathbf{C}\mathbf{a}_1)'\left[\mathbf{C}\mathbf{E}_1\mathbf{C}'\right]^+(\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{a}_1).\end{aligned}$$

Proof of (g). From Lemma 3.3, we see that

$$\begin{aligned}&\mathbf{F}^+\mathbf{E}_i\mathbf{C}'\left[\mathbf{C}\mathbf{E}_i\mathbf{C}' - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+'}\left(\mathbf{F}^+\mathbf{E}_i\mathbf{F}^{+'}\right)^+\mathbf{F}^+\mathbf{E}_i\mathbf{C}'\right]^+\mathbf{C}\mathbf{a}_i \\ &\quad = \left[\mathbf{F}^+\mathbf{E}_i\left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)\mathbf{R}'\right]\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'\right]^+\mathbf{C}\mathbf{a}_i \\ &\quad = \left[\mathbf{F}^+\left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)\mathbf{E}_i\mathbf{R}'\right]\left[\mathbf{C}\mathbf{E}_i\mathbf{C}'\right]^+\mathbf{C}\mathbf{a}_i \\ &\quad = \mathbf{0}.\end{aligned}$$

Proof of (h). From Lemma 3.3, we have

$$\mathbf{C}\mathbf{a}_i - \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime}\right)^{+}\mathbf{F}^{+}\mathbf{a}_i = \mathbf{C}\mathbf{a}_1.$$

Lemma 3.5. Let $\mathbf{F} \in \mathbb{R}_{p\times q}$ and $\mathbf{C} = \mathbf{R}\left[\mathbf{I} - \mathbf{F}\mathbf{F}^{+}\right]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank\,(\mathbf{C}) = p - q$, and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\,(\mathbf{E}_i) < p$ for $i = 1, 2, ..., m$, such that properties (1) - (3) hold. Also, let

$$\mathbf{A} = \begin{bmatrix} \mathbf{F}^{+}\mathbf{E}_i\mathbf{F}^{+\prime} & \mathbf{F}^{+}\mathbf{E}_i\mathbf{C}^{\prime} \\ \mathbf{C}\mathbf{E}_i\mathbf{F}^{+\prime} & \mathbf{C}\mathbf{E}_i\mathbf{C}^{\prime} \end{bmatrix}.$$

Then,

$$\mathbf{A}^{+} = \begin{bmatrix} \mathbf{F}^{\prime}\mathbf{E}_i^{+}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{+} \end{bmatrix}.$$

Proof: The proof is a direct result of Lemmas 3.1.d, 3.2, and 3.3.

Lemma 3.6. Let $\mathbf{M}$ be defined as $\mathbf{M} = \mathbf{F}\mathbf{G}$ with $rank\,(\mathbf{F}) = rank\,(\mathbf{G}) = q$, where $\mathbf{a}_i \in \mathbb{R}_{p\times 1}$, $\mathbf{F} \in \mathbb{R}_{p\times q}$, $\mathbf{G} \in \mathbb{R}_{q\times(m-1)(p+1)}$, and $\mathbf{E}_i \in \mathbb{R}^{S}_{p\times p}$ with $rank\,(\mathbf{E}_i) < p$, and let $\mathbf{C} = \mathbf{R}\left[\mathbf{I} - \mathbf{F}\mathbf{F}^{+}\right]$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank\,(\mathbf{C}) = p - q$. Also, let $rank\,(\mathbf{M}) < q$, where $1 \leq q < p$, and let $\mathbf{F}$ and $\mathbf{G}$ be matrix components of a full rank decomposition of $\mathbf{M}$ with $rank\,(\mathbf{F}) = rank\,(\mathbf{G}) = q$. Then, for $i = 1, 2, ..., m$,

$$(\mathbf{H}\mathbf{a}_i)^{\prime}\,\mathbf{H}\mathbf{E}_i^{+}\mathbf{H}^{\prime}\,(\mathbf{H}\mathbf{a}_i) = \mathbf{a}_i\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\mathbf{E}_i^{+}\mathbf{F}\right)\mathbf{F}^{+}\mathbf{a}_i + \mathbf{a}_1^{\prime}\mathbf{C}^{\prime}\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{+}\mathbf{C}\mathbf{a}_1.$$

Proof: We have that

$$\begin{aligned}
(\mathbf{H}\mathbf{a}_i)^{\prime}\left(\mathbf{H}\mathbf{E}_i^{+}\mathbf{H}^{\prime}\right)(\mathbf{H}\mathbf{a}_i) &= \mathbf{a}_i^{\prime}\mathbf{H}^{\prime}\begin{bmatrix} \mathbf{F}^{\prime}\mathbf{E}_i^{+}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{+} \end{bmatrix}\mathbf{H}\mathbf{a}_i \\
&= \begin{bmatrix} \mathbf{a}_i^{\prime}\mathbf{F}^{+\prime} & \mathbf{a}_i^{\prime}\mathbf{C}^{\prime} \end{bmatrix}\begin{bmatrix} \mathbf{F}^{\prime}\mathbf{E}_i^{+}\mathbf{F} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{+} \end{bmatrix}\begin{bmatrix} \mathbf{F}^{+}\mathbf{a}_i \\ \mathbf{C}\mathbf{a}_i \end{bmatrix} \\
&= \mathbf{a}_i^{\prime}\mathbf{F}^{+\prime}\mathbf{F}^{\prime}\mathbf{E}_i^{+}\mathbf{F}\mathbf{F}^{+}\mathbf{a}_i + \mathbf{a}_1^{\prime}\mathbf{C}^{\prime}\left(\mathbf{C}\mathbf{E}_1\mathbf{C}^{\prime}\right)^{+}\mathbf{C}\mathbf{a}_1.
\end{aligned}$$

Lemma 3.7. Let $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$ for $i \in \{1, 2, ..., m\}$, and let $\mathbf{H}$ be the full-rank matrix $\mathbf{H} \equiv [\mathbf{F}^{+\prime} \ \ \mathbf{C}]'$, where $\mathbf{F}$ and $\mathbf{C}$ are defined previously. Then, the conditional density function of $\mathbf{F}^{+}\mathbf{x}|\mathbf{C}\mathbf{x}$ is

$$g\left(\mathbf{F}^{+}\mathbf{x}|\mathbf{C}\mathbf{x}\right) = \frac{1}{\Phi\left(\beta_0\right)}\varphi_q^S\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_i, \mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \times \Phi\left(l_{0i} + \mathbf{l}_{1i}'\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_i\right)\right), \tag{3.13}$$

where

$$\beta_0 \equiv \frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}_1\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - \left(\mathbf{C}\boldsymbol{\gamma}_1\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\left(\mathbf{C}\boldsymbol{\gamma}_1\right)}}, \tag{3.14}$$

$$l_{0i} \equiv \frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}_1\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1 - k_i}},$$

and

$$\mathbf{l}_{1i} \equiv \frac{\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\boldsymbol{\gamma}_i\right)}{\sqrt{1 - k_i}}$$

for $k_i = \boldsymbol{\gamma}_i\mathbf{F}^{+\prime}\mathbf{F}'\boldsymbol{\Sigma}_i^{+}\mathbf{F}\mathbf{F}^{+}\boldsymbol{\gamma}_i + \boldsymbol{\gamma}_1'\mathbf{C}'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\mathbf{C}\boldsymbol{\gamma}_1$.

Proof: Using Proposition 3.4, we have

$$l_{0i} = \frac{\delta_0 + \left(\mathbf{C}\boldsymbol{\gamma}_1\right)'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)\left(\mathbf{C}\boldsymbol{\gamma}_1\right)}{\sqrt{1 - k_i}}.$$

In addition, from Proposition 3.4, the $MSSN$ location parameter is

$$\mathbf{F}^{+}\boldsymbol{\mu}_i + \mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{C}'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)\left(\mathbf{C}\mathbf{x} - \mathbf{C}\boldsymbol{\mu}_1\right) = \mathbf{F}^{+}\boldsymbol{\mu}_i$$

because $\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{C}'\right)' = \mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} = \mathbf{0}$ by Lemma 3.3. Also, the dispersion parameter is

$$\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} - \mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{C}'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} = \mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}$$

because $\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} = \mathbf{0}$ by Lemma 3.3. Finally, Proposition 3.4 gives

$$\mathbf{l}_{1i} = \frac{\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} - \mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{C}'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\mathbf{C}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\boldsymbol{\gamma}_i - \mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{C}'\left(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)^{+}\mathbf{C}\boldsymbol{\gamma}_1\right)}{\sqrt{1 - k_i}},$$

which, because $\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{C}' = \mathbf{0}$, simplifies to

$$\mathbf{l}_{1i} = \frac{\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\boldsymbol{\gamma}_i\right)}{\sqrt{1 - k_i}}.$$

Remark 3.1. We note that the conditional density function in (3.13) is not a $MSSN$ density function as defined in (3.5).

Lemma 3.8. Define the random vector $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$ for class Π_i, where $i \in \{1, 2, ..., m\}$, and let $\mathbf{H}$ be the full-rank matrix $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$ and $\mathbf{F}$ and $\mathbf{C}$ are defined previously. Then, $\mathbf{Cx} \sim SSN_{p-q}(\mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}', \delta_0, \mathbf{C}\boldsymbol{\gamma}_1)$, where

$$f(\mathbf{Cx}|\Pi_i) = \frac{1}{\Phi(\delta_0)} \varphi_{p-q}^S(\mathbf{Cx}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}') \Phi\left(\frac{\delta_0 + (\mathbf{C}\boldsymbol{\gamma}_1)'(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}')^+(\mathbf{Cx} - \mathbf{C}\boldsymbol{\mu}_1)}{\sqrt{1 - (\mathbf{C}\boldsymbol{\gamma}_1)'(\mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}')^+(\mathbf{C}\boldsymbol{\gamma}_1)}} \right)$$

for $i \in \{1, ..., m\}$.

Proof: The proof of the lemma follows from (ii) of Proposition 3.3.

Lemma 3.9. Let $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, where $\mathbf{H} \in \mathbb{R}_{p \times p}$ with $rank(\mathbf{H}) = p$. Then, $p(\mathbf{Hx}|\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$, where $p(\cdot|\boldsymbol{\theta})$ is the p-dimensional $MSSN$ density function with parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma}\}$.

Proof: We have that

$$\begin{aligned}
p(\mathbf{Hx}|\boldsymbol{\theta}) &= \frac{1}{\Phi(\delta_0)} \varphi_p^S(\mathbf{Hx}; \mathbf{H}\boldsymbol{\mu}, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}') \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')^+(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')^+\mathbf{H}\boldsymbol{\gamma}}} \right) \\
&= \frac{1}{\Phi(\delta_0)} \frac{1}{(2\pi)^{n/2}|\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu})'(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}')^+(\mathbf{Hx} - \mathbf{H}\boldsymbol{\mu}) \right\} \times \\
&\quad \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\mathbf{H}'\mathbf{H}'^{-1}\boldsymbol{\Sigma}^+\mathbf{H}^{-1}\mathbf{H}(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}'\mathbf{H}'\mathbf{H}'^{-1}\boldsymbol{\Sigma}^+\mathbf{H}^{-1}\mathbf{H}\boldsymbol{\gamma}}} \right) \\
&= \frac{1}{\Phi(\delta_0)} \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{H}'\mathbf{H}'^{-1}\boldsymbol{\Sigma}^+\mathbf{H}^{-1}\mathbf{H}(\mathbf{x} - \boldsymbol{\mu}) \right\} \times \\
&\quad \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^+(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}} \right) \\
&= \frac{1}{\Phi(\delta_0)} \varphi_p^S(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \Phi\left(\frac{\delta_0 + \boldsymbol{\gamma}'\boldsymbol{\Sigma}^+(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{1 - \boldsymbol{\gamma}'\boldsymbol{\Sigma}^+\boldsymbol{\gamma}}} \right) \\
&= p(\mathbf{x}|\boldsymbol{\theta}).
\end{aligned}$$

Theorems 2.2 through 2.5 are important in that if their conditions hold, we obtain a linear dimension-compression matrix for the reduced q-dimensional subspace such that the $BPMC$ in the q-dimensional space is equal to the $BPMC$ for

the original p-dimensional feature space. In other words, provided the conditions of Theorems 2.2 through 2.5 hold, we have that the linear feature-reduction matrix $\mathbf{F}^{+} \in \mathbb{R}_{q\times p}$ exists such that $BPMC(p) = BPMC(q)$, where $rank(\mathbf{M}) = q < p$.

In the next theorem, we determine the conditions for the existence of a linear data-compression matrix when the populations are $MSSN$ with certain conditions on the dispersion parameters.

Theorem 3.3. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by distribution $SSN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$ for some $k \in \{2, ..., m\}$, dispersion parameter $\boldsymbol{\Sigma}$ with $rank(\boldsymbol{\Sigma}) = k < p$ such that $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_1) \in \mathcal{C}(\boldsymbol{\Sigma})$, skew parameter $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \in \mathcal{N}(\mathbf{C})$, and scalar $\delta_{0i} = \delta_0$, $i = 1, 2, ..., m$. Also, let

$$\mathbf{M}_S \equiv \left[\boldsymbol{\Sigma}^{+}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) | \boldsymbol{\Sigma}^{+}(\boldsymbol{\mu}_3 - \boldsymbol{\mu}_1) | ... | \boldsymbol{\Sigma}^{+}(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1)\right],$$

where $\mathbf{M}_S = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}_S$ with $rank(\mathbf{M}_S) = q < p$. Then, the p-variate Bayes classifier assigns the unlabeled observation vector $\mathbf{x}$ to Π_k if and only if the q-variate Bayes classifier assigns $\mathbf{F}^{+}\mathbf{x}$ to Π_k for $k \in \{1, 2, ..., m\}$.

Proof: Let

$$\mathbf{w} = \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^{+}_{q\times p} \\ \mathbf{C}_{(p-q)\times p} \end{bmatrix} \mathbf{x},$$

where $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}, \delta_0, \boldsymbol{\gamma})$, $i = 1, 2, ..., m$. Let $\mathbf{H}$ be the full-rank matrix $\mathbf{H} \equiv \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$ with $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^{+}\right)$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{R}) = p - q$. Then, $\mathbf{w} \sim SSN_p(\mathbf{H}\boldsymbol{\mu}_i, \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}', \delta_0, \mathbf{H}\boldsymbol{\gamma})$ such that

$$\mathbf{H}\boldsymbol{\mu}_i = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\mu}_i \\ \mathbf{C}\boldsymbol{\mu}_i \end{bmatrix} \text{ and } \mathbf{H}\boldsymbol{\Sigma}\mathbf{H}' = \begin{bmatrix} \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' \end{bmatrix} \text{ for } i \in \{1, 2, ..., m\}.$$

By Lemma 3.4,

$$g(\mathbf{u}|\mathbf{y}) = \frac{1}{\Phi(\delta_0)} \varphi_q^S\left(\mathbf{F}^{+}\mathbf{x}; \mathbf{F}^{+}\boldsymbol{\mu}_i, \mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(\delta_0 + \mathbf{l}_1' \left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+} \left(\mathbf{F}^{+}\mathbf{x} - \mathbf{F}^{+}\boldsymbol{\mu}_i\right)\right),$$

where

$$\mathbf{l}_1 = \frac{\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\mathbf{F}^{+}\boldsymbol{\gamma}}{\sqrt{1-\boldsymbol{\gamma}^{\prime}\mathbf{F}^{+\prime}\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{+}\mathbf{F}^{+}\boldsymbol{\gamma}}}.$$

Also, by Proposition 3.2, the marginal density of $\mathbf{y}$ is the $MSSN$ density

$$h\left(\mathbf{y}\right) = \varphi_{p-q}^{S}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_1,\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)\Phi\left(\frac{\delta_0+\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right).$$

The p-variate Bayes classification procedure assigns $\mathbf{x}$ to Π_j if and only if

$$\alpha_j\frac{1}{\Phi\left(\delta_0\right)}\varphi_p^{S}\left(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}\right)\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}^{\prime}\boldsymbol{\Sigma}^{+}\left(\mathbf{x}-\boldsymbol{\mu}_j\right)}{\sqrt{1-\boldsymbol{\gamma}^{\prime}\boldsymbol{\Sigma}^{+}\boldsymbol{\gamma}}}\right) >$$
$$\alpha_j\frac{1}{\Phi\left(\delta_0\right)}\varphi_p^{S}\left(\mathbf{x};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}\right)\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}^{\prime}\boldsymbol{\Sigma}^{+}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)}{\sqrt{1-\boldsymbol{\gamma}^{\prime}\boldsymbol{\Sigma}^{+}\boldsymbol{\gamma}}}\right),$$

$i=1,2,...,m$, which is equivalent to

$$\alpha_j\frac{1}{\Phi\left(\delta_0\right)}\varphi_p^{S}\left(\mathbf{H}\mathbf{x};\mathbf{H}\boldsymbol{\mu}_j,\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\prime}\right)\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}^{\prime}\mathbf{H}^{\prime}\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\prime}\right)^{+}\left(\mathbf{H}\mathbf{x}-\mathbf{H}\boldsymbol{\mu}_j\right)}{\sqrt{\boldsymbol{\gamma}^{\prime}\mathbf{H}^{\prime}\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\prime}\right)^{+}\mathbf{H}\boldsymbol{\gamma}}}\right) >$$
$$\alpha_i\frac{1}{\Phi\left(\delta_0\right)}\varphi_p^{S}\left(\mathbf{H}\mathbf{x};\mathbf{H}\boldsymbol{\mu}_i,\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\prime}\right)\Phi\left(\frac{\delta_0+\boldsymbol{\gamma}^{\prime}\mathbf{H}^{\prime}\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\prime}\right)^{+}\left(\mathbf{H}\mathbf{x}-\mathbf{H}\boldsymbol{\mu}_i\right)}{\sqrt{\boldsymbol{\gamma}^{\prime}\mathbf{H}^{\prime}\left(\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}^{\prime}\right)^{+}\mathbf{H}\boldsymbol{\gamma}}}\right)$$

or

$$\alpha_j\frac{1}{\Phi\left(\delta_0\right)}\varphi_q^{S}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_j,\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)\Phi\left(l_0+\mathbf{l}_1^{\prime}\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_j\right)\right)\times$$
$$\varphi_{p-q}^{S}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_1,\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)\Phi\left(\frac{\delta_0+\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right) >$$

$$\alpha_i\frac{1}{\Phi\left(\delta_0\right)}\varphi_q^{S}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_i,\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)\Phi\left(l_0+\mathbf{l}_1^{\prime}\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_i\right)\right)\times$$
$$\varphi_{p-q}^{S}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_1,\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)\Phi\left(\frac{\delta_0+\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)$$

by Lemma 3.7. Because the marginal density

$$\varphi_{p-q}^{S}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_1,\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)\Phi\left(\frac{\delta_0+\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\mathbf{x}-\mathbf{C}\boldsymbol{\mu}_1\right)}{\sqrt{1-\left(\mathbf{C}\boldsymbol{\gamma}\right)^{\prime}\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^{\prime}\right)^{+}\left(\mathbf{C}\boldsymbol{\gamma}\right)}}\right)$$

does not depend on k for $k = 2, ..., m$, we have that

$$\alpha_j \frac{1}{\Phi(\delta_0)} \varphi_q^S \left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_j, \mathbf{F}^+\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_0 + \mathbf{l}_1' \left[\left(\mathbf{F}^+\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^+ \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_j\right)\right) >$$
$$\alpha_i \frac{1}{\Phi(\delta_0)} \varphi_q^S \left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right) \Phi\left(l_0 + \mathbf{l}_1' \left[\left(\mathbf{F}^+\boldsymbol{\Sigma}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^+ \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right).$$

Hence, if the Bayes classifier classifies $\mathbf{x}$ into Π_j, then $\mathbf{F}^+\mathbf{x}$ is classified into Π_j. The preceding arguments are reversible, thus completing the proof of the equivalence of the original p-variate Bayes classification and the transformed q-variate Bayes classification procedure.

Theorem 3.4. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by distribution $SSN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$, dispersion parameter $\boldsymbol{\Sigma}_i$ with $rank(\boldsymbol{\Sigma}_i) = k < p$ such that $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_k$ and $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_1) \in \mathcal{C}(\boldsymbol{\Sigma}_i)$, skew parameter $\boldsymbol{\gamma}_i$ such that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_k$, and scalar δ_0 for $i \in \{1, 2, ..., m\}$. Next, let

$$\mathbf{M}_S \equiv \left[\boldsymbol{\Sigma}_2^+(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) | ... | \boldsymbol{\Sigma}_m^+(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1) | \boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1 | ... | \boldsymbol{\Sigma}_m - \boldsymbol{\Sigma}_1 | \boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1 | ... | \boldsymbol{\gamma}_m - \boldsymbol{\gamma}_1\right],$$

where $\mathbf{M} = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}_S$ with $rank(\mathbf{M}_S) = q < p$, and let $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$. Then, the p-variate Bayes classification procedure assigns the unlabeled observation vector $\mathbf{x}$ to Π_k if and only if the q-dimensional Bayes classification procedure assigns $\mathbf{F}^+\mathbf{x}$ to Π_k for $k \in \{1, 2, ..., m\}$.

Proof: Let $\mathbf{w} = \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^+_{q\times p} \\ \mathbf{C}_{(p-q)\times p} \end{bmatrix} \mathbf{x}$, where $\mathbf{x} \sim SSN_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i)$, $i = 1, 2, ..., m$. Let $\mathbf{H}$ be a full-rank $p \times p$ matrix defined by $\mathbf{H} = \begin{bmatrix} \mathbf{F}^{+\prime} & \mathbf{C}' \end{bmatrix}'$ with $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{FF}^+\right)$, where $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$ such that $rank(\mathbf{R}) = p - q$. Then, $\mathbf{w} \sim SSN_p(\mathbf{H}\boldsymbol{\mu}_i, \mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}', \delta_0, \mathbf{H}\boldsymbol{\gamma}_i)$ such that

$$\mathbf{H}\boldsymbol{\mu}_i = \begin{bmatrix} \mathbf{F}^+\boldsymbol{\mu}_i \\ \mathbf{C}\boldsymbol{\mu}_i \end{bmatrix} \text{ and } \mathbf{H}\boldsymbol{\Sigma}_i\mathbf{H}' = \begin{bmatrix} \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}' \end{bmatrix} \text{ for } i = 1, 2, ..., m.$$

By Lemma 3.4,

$$p\left(\mathbf{u}|\mathbf{y}\right)=\frac{1}{\Phi\left(l_{0i}\right)}\varphi_{q}^{S}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_{i},\mathbf{F}^{+}\boldsymbol{\Sigma}_{i}\mathbf{F}^{+\prime}\right)\Phi\left(l_{0i}+\mathbf{l}_{1i}^{\prime}\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_{i}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_{i}\right)\right),$$

where

$$\mathbf{l}_{1i}=\frac{\left(\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_{i}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right)^{+}\mathbf{F}^{+}\boldsymbol{\gamma}_{i}}{\sqrt{1-\boldsymbol{\gamma}_{i}^{\prime}\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\boldsymbol{\Sigma}_{i}^{+}\mathbf{F}\right)^{+}\mathbf{F}^{+}\boldsymbol{\gamma}_{i}}}$$

and

$$l_{0i}=\frac{\delta_{0i}}{\sqrt{1-\boldsymbol{\gamma}_{i}^{\prime}\mathbf{F}^{+\prime}\left(\mathbf{F}^{\prime}\boldsymbol{\Sigma}_{i}^{+}\mathbf{F}\right)^{+}\mathbf{F}^{+}\boldsymbol{\gamma}_{i}}}.$$

Also, because $\boldsymbol{\gamma}_{1}\in\mathcal{N}\left(\mathbf{C}\right)$, the marginal density of $\mathbf{y}$ is

$$h\left(\mathbf{u}\right)=\varphi_{p-q}^{S}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_{1},\mathbf{C}\boldsymbol{\Sigma}_{1}\mathbf{C}^{\prime}\right).$$

Let $p(\cdot|\Pi_{i})$ denote the p-dimensional $MSSN$ density corresponding to population Π_{i}, $i=1,2,...,m$. Recall that the p-variate Bayes classification procedure assigns $\mathbf{x}$ to Π_{j} if and only if

$$\alpha_{j}\frac{1}{\Phi\left(\delta_{0j}\right)}\varphi_{p}^{S}\left(\mathbf{x};\boldsymbol{\mu}_{j},\boldsymbol{\Sigma}_{j}\right)\Phi\left(\frac{\delta_{0j}+\boldsymbol{\gamma}_{j}^{\prime}\ \boldsymbol{\Sigma}_{j}^{+}\left(\mathbf{x}-\boldsymbol{\mu}_{j}\right)}{\sqrt{1-\boldsymbol{\gamma}_{j}^{\prime}\ \boldsymbol{\Sigma}_{j}^{+}\boldsymbol{\gamma}_{j}}}\right)>$$
$$\alpha_{i}\frac{1}{\Phi\left(\delta_{0i}\right)}\varphi_{p}^{S}\left(\mathbf{x};\boldsymbol{\mu}_{i},\boldsymbol{\Sigma}_{i}\right)\Phi\left(\frac{\delta_{0i}+\boldsymbol{\gamma}_{i}^{\prime}\ \boldsymbol{\Sigma}_{i}^{+}\left(\mathbf{x}-\boldsymbol{\mu}_{i}\right)}{\sqrt{1-\boldsymbol{\gamma}_{i}^{\prime}\ \boldsymbol{\Sigma}_{i}^{+}\boldsymbol{\gamma}_{i}}}\right),$$

$i=1,2,...,m$, $i\neq j$, which is equivalent to

$$\alpha_{j}\frac{1}{\Phi\left(\delta_{0j}\right)}\varphi_{p}^{S}\left(\mathbf{H}\mathbf{x};\mathbf{H}\boldsymbol{\mu}_{j},\mathbf{H}\boldsymbol{\Sigma}_{j}\mathbf{H}^{\prime}\right)\Phi\left(\frac{\delta_{0j}+\boldsymbol{\gamma}_{j}^{\prime}\mathbf{H}^{\prime}\ \left(\mathbf{H}\boldsymbol{\Sigma}_{j}\mathbf{H}^{\prime}\right)^{+}\left(\mathbf{H}\mathbf{x}-\mathbf{H}\boldsymbol{\mu}_{j}\right)}{\sqrt{1-\boldsymbol{\gamma}_{j}^{\prime}\mathbf{H}^{\prime}\ \left(\mathbf{H}\boldsymbol{\Sigma}_{j}\mathbf{H}^{\prime}\right)^{+}\boldsymbol{H}\boldsymbol{\gamma}_{j}}}\right)>$$
$$\alpha_{i}\frac{1}{\Phi\left(\delta_{0i}\right)}\varphi_{p}^{S}\left(\mathbf{H}\mathbf{x};\mathbf{H}\boldsymbol{\mu}_{i},\mathbf{H}\boldsymbol{\Sigma}_{i}\mathbf{H}^{\prime}\right)\Phi\left(\frac{\delta_{0i}+\boldsymbol{\gamma}_{i}^{\prime}\mathbf{H}^{\prime}\ \left(\mathbf{H}\boldsymbol{\Sigma}_{i}\mathbf{H}^{\prime}\right)^{+}\left(\mathbf{H}\mathbf{x}-\mathbf{H}\boldsymbol{\mu}_{i}\right)}{\sqrt{1-\boldsymbol{\gamma}_{i}^{\prime}\mathbf{H}^{\prime}\ \left(\mathbf{H}\boldsymbol{\Sigma}_{i}\mathbf{H}^{\prime}\right)^{+}\mathbf{H}\boldsymbol{\gamma}_{i}}}\right).$$

Therefore,

$$\alpha_{j}\frac{1}{\Phi\left(l_{0j}\right)}\varphi_{q}^{S}\left(\mathbf{F}^{+}\mathbf{x};\mathbf{F}^{+}\boldsymbol{\mu}_{j},\mathbf{F}^{+}\boldsymbol{\Sigma}_{j}\mathbf{F}^{+\prime}\right)\Phi\left(l_{0j}+\mathbf{l}_{1j}^{\prime}\left[\left(\mathbf{F}^{+}\boldsymbol{\Sigma}_{j}\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^{+}\left(\mathbf{F}^{+}\mathbf{x}-\mathbf{F}^{+}\boldsymbol{\mu}_{j}\right)\right)\times$$
$$\varphi_{p-q}^{S}\left(\mathbf{C}\mathbf{x};\mathbf{C}\boldsymbol{\mu}_{1},\mathbf{C}\boldsymbol{\Sigma}_{1}\mathbf{C}^{\prime}\right)>$$

$$\alpha_i \frac{1}{\Phi\left(l_{0i}\right)} \varphi_q^S\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}_{1i}' \left[\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^+ \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right) \times$$
$$\varphi_{p-q}^S\left(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)$$

by Lemma 3.7. Hence,

$$\alpha_j \frac{1}{\Phi\left(l_{0j}\right)} \varphi_q^S\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_j, \mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right) \Phi\left(l_{0j} + \mathbf{l}_{1j}' \left[\left(\mathbf{F}^+\boldsymbol{\Sigma}_j\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^+ \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_j\right)\right) >$$
$$\alpha_i \frac{1}{\Phi\left(l_{0i}\right)} \varphi_q^S\left(\mathbf{F}^+\mathbf{x}; \mathbf{F}^+\boldsymbol{\mu}_i, \mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right) \Phi\left(l_{0i} + \mathbf{l}_{1i}' \left[\left(\mathbf{F}^+\boldsymbol{\Sigma}_i\mathbf{F}^{+\prime}\right)^{\frac{1}{2}}\right]^+ \left(\mathbf{F}^+\mathbf{x} - \mathbf{F}^+\boldsymbol{\mu}_i\right)\right)$$

because

$$\varphi_{p-q}^S\left(\mathbf{C}\mathbf{x}; \mathbf{C}\boldsymbol{\mu}_1, \mathbf{C}\boldsymbol{\Sigma}_1\mathbf{C}'\right)$$

does not depend on k for $k = 2, ..., m$ and $\boldsymbol{\gamma}_1 \in \mathcal{N}\left(\mathbf{C}\right)$. Thus, if the p-dimensional Bayes classifier classifies $\mathbf{x}$ into Π_j, then the reduced q-dimensional Bayes classifier classifies $\mathbf{F}^+\mathbf{x}$ into Π_j. The preceding arguments are reversible, therefore completing the proof of the equivalence of the original p-variate Bayes classification and the transformed q-variate Bayes classification procedure.

Corollary 3.1. Let Π_i have *a priori* probability $\alpha_i > 0$ and be represented by distribution $SSN_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \delta_0, \boldsymbol{\gamma}_i\right)$ with location parameter $\boldsymbol{\mu}_i$ such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_k$, dispersion parameter $\boldsymbol{\Sigma}_i$ with $rank\left(\boldsymbol{\Sigma}_i\right) = k < p$, where $\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right) \in \mathcal{C}\left(\boldsymbol{\Sigma}_i\right)$ for some $k \in \{2, 3, ..., m\}$ and $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_k$, skew parameter $\boldsymbol{\gamma}_i$ such that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\gamma}_k$ for some $k \in \{2, ..., m\}$, and scalar δ_0 for $i \in \{1, 2, ..., m\}$. In addition, let Π_1 be represented by the singular normal distribution $SN_p\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right)$, where $rank\left(\boldsymbol{\Sigma}_1\right) = k < p$. Next, let

$$\mathbf{M}_S \equiv \left[\boldsymbol{\Sigma}_2^+\left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\right)|...|\boldsymbol{\Sigma}_m^+\left(\boldsymbol{\mu}_m - \boldsymbol{\mu}_1\right)|\boldsymbol{\Sigma}_2 - \boldsymbol{\Sigma}_1|...|\boldsymbol{\Sigma}_m - \boldsymbol{\Sigma}_1|\boldsymbol{\gamma}_2|...|\boldsymbol{\gamma}_m\right],$$

where $\mathbf{M}_S = \mathbf{FG}$ is a full-rank decomposition of $\mathbf{M}_S$ with $rank\left(\mathbf{M}_S\right) = q < p$, and let $\boldsymbol{\gamma}_1 \in \mathcal{N}\left(\mathbf{C}\right)$, where $\mathbf{C} = \mathbf{R}\left(\mathbf{I} - \mathbf{F}\mathbf{F}^+\right)$ and $\mathbf{R} \in \mathbb{R}_{(p-q)\times p}$. Also, let $\mathbf{x} \in \mathbb{R}_{p\times 1}$ be an unlabeled observation vector belonging to Π_j for $j \in \{1, 2, ..., m\}$. Then, the

p-variate Bayes classifier assigns $\mathbf{x}$ to Π_k for $k \in \{1, 2, ..., m\}$ if and only if the q-dimensional Bayes classifier assigns $\mathbf{F}^{+}\mathbf{x}$ to Π_k.

Proof: The proof follows immediately from Theorem 3.4 because $\boldsymbol{\gamma}_1 = \mathbf{0}$.

3.3 Examples

Example 3.1. In the first example, we demonstrate a low-dimensional representation for two $MSSN$ populations having unequal location, dispersion, and skew parameters but equal scalar parameters. Consider the two seven-dimensional populations $SSN_7(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \delta_0, \boldsymbol{\gamma}_1)$ and $SSN_7(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \delta_0, \boldsymbol{\gamma}_2)$, where

$$\Pi_1 : \boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 3.14 & .25 & .25 & .25 & .25 & .25 & .25 \\ .25 & 3.14 & .25 & .25 & .25 & .25 & .25 \\ .25 & .25 & 3.14 & .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & 3.14 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 & 3.14 & .25 & .25 \\ .25 & .25 & .25 & .25 & .25 & .25 & .25 \\ .25 & .25 & .25 & .25 & .25 & .25 & .25 \end{bmatrix}, \boldsymbol{\gamma}_1 = \begin{bmatrix} .354 \\ .364 \\ .374 \\ .384 \\ .394 \\ .409 \\ .409 \end{bmatrix}$$

$$\Pi_2 : \boldsymbol{\mu}_2 = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 3 & .11 & .11 & .11 & .11 & .11 & .11 \\ .11 & 3 & .11 & .11 & .11 & .11 & .11 \\ .11 & .11 & 3 & .11 & .11 & .11 & .11 \\ .11 & .11 & .11 & 3 & .11 & .11 & .11 \\ .11 & .11 & .11 & .11 & 3 & .11 & .11 \\ .11 & .11 & .11 & .11 & .11 & .11 & .11 \\ .11 & .11 & .11 & .11 & .11 & .11 & .11 \end{bmatrix}, \boldsymbol{\gamma}_2 = \begin{bmatrix} .264 \\ .274 \\ .284 \\ .294 \\ .304 \\ .319 \\ .320 \end{bmatrix}$$

with $\delta_{0i} = 0$, $i = 1, 2$. Using Theorem 3.4 to formulate a dimension-reduction matrix, we obtain

$$\mathbf{M}_S = \begin{bmatrix} -2.08 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \\ -1.73 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \\ -.69 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \\ 0 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \\ -1.04 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \\ 39.13 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \\ 39.13 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.14 & -.09 \end{bmatrix},$$

where $rank\,(\mathbf{M}_S) = 2$. Therefore, by Theorem 3.4, the original seven-dimensional density functions can be transformed to the reduced dimension $q = 2$ without increasing the probability of misclassification under the Bayes assignment. An optimal two-dimensional representation space is the column space of $\mathbf{F}$, where

$$\mathbf{F}' = \begin{bmatrix} .445 & .445 & .445 & .445 & .445 & .075 & .075 \\ .549 & .546 & -.374 & -.326 & -.394 & 0 & 0 \end{bmatrix}'.$$

We now have the reduced parameters

$$\Pi_1 : \boldsymbol{\mu}_{1R} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{1R} = \begin{bmatrix} 4.27 & 0 \\ 0 & 2.89 \end{bmatrix}, \boldsymbol{\gamma}_{1R} = \begin{bmatrix} .892 \\ -.027 \end{bmatrix}, \text{ and}$$

$$\Pi_2 : \boldsymbol{\mu}_{2R} = \begin{bmatrix} 4.75 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma}_{2R} = \begin{bmatrix} 3.48 & 0 \\ 0 & 2.89 \end{bmatrix}, \boldsymbol{\gamma}_{2R} = \begin{bmatrix} .679 \\ -.027 \end{bmatrix}.$$

Immediately above, we portray a two-dimensional representation of the original seven-dimensional *MSSN* populations. By using the *SVD*, we can obtain a one-dimensional representation of both populations; however, we lose some discriminatory information by applying this procedure. For the first population, we obtain the parameters

$$\Pi_1 : \mu_1 = 0, \sigma_1 = 4.27, \text{ and } \gamma_1 = .892.$$

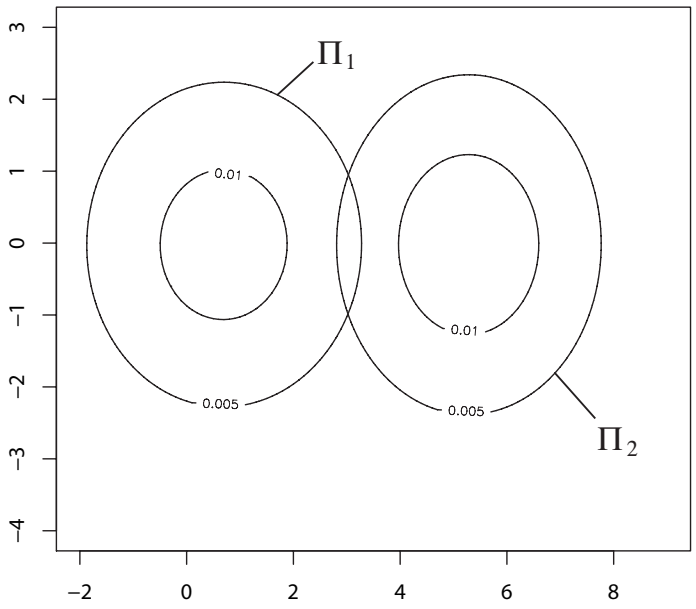

Figure 3.1: The optimal two-dimensional representation for both seven-dimensional singular skew-normal populations in Example 3.1.

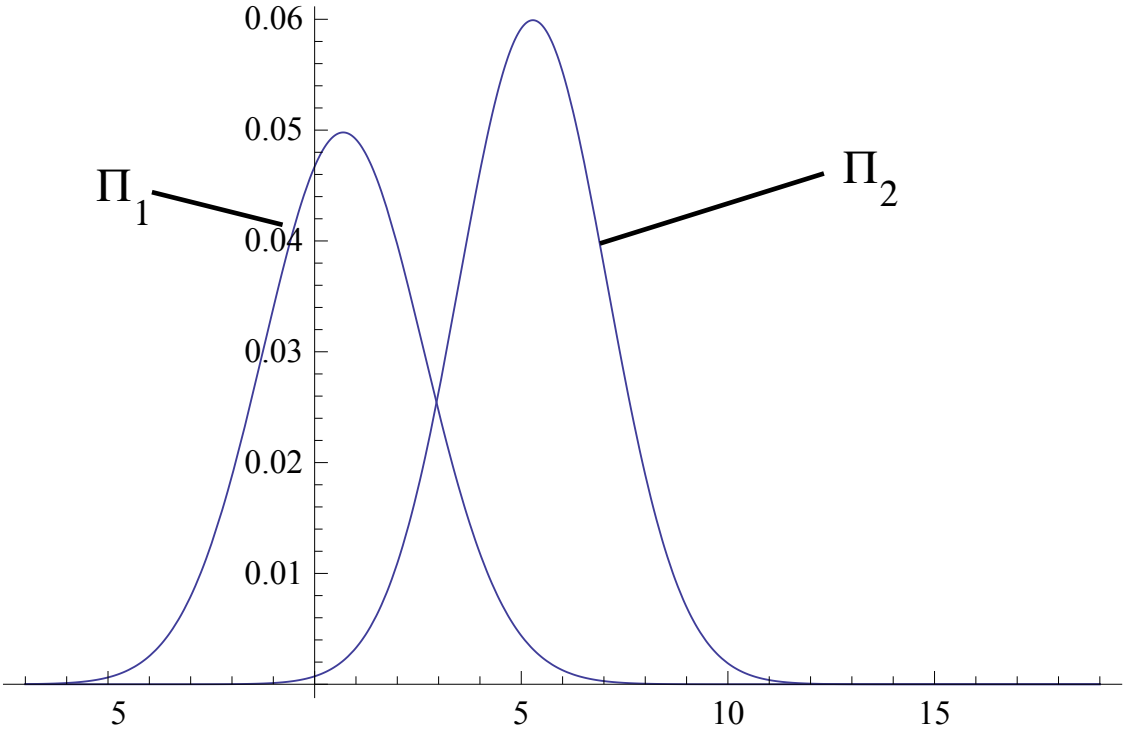

Figure 3.2: A one-dimensional approximate representation for both seven-dimensional singular skew-normal populations in Example 3.1.

For the second population, we have parameters

$$\Pi_2 : \mu_2 = 11.87, \sigma_2 = 3.48, \text{ and } \gamma_2 = .679.$$

Example 3.2. In the second example, we demonstrate a low-dimensional representation for three $MSSN$ populations having unequal location, dispersion, and skew parameters. Consider the three six-dimensional populations $SSN_6(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \delta_{01}, \boldsymbol{\gamma}_1)$, $SSN_6(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \delta_{02}, \boldsymbol{\gamma}_2)$, and $SSN_6(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3, \delta_{03}, \boldsymbol{\gamma}_3)$, with

$$\Pi_1 : \boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & .7 & .7 & .7 & .7 & .7 \\ .7 & 2 & .7 & .7 & .7 & .7 \\ .7 & .7 & 2 & .7 & .7 & .7 \\ .7 & .7 & .7 & 2 & .7 & .7 \\ .7 & .7 & .7 & .7 & .7 & .7 \\ .7 & .7 & .7 & .7 & .7 & .7 \end{bmatrix}, \boldsymbol{\gamma}_1 = \begin{bmatrix} .763 \\ .698 \\ .698 \\ .821 \\ .829 \\ .745 \end{bmatrix},$$

$$\Pi_2 : \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ 4 \\ 4 \\ 2 \\ 2 \\ 3 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2.3 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2.3 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2.3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2.3 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\gamma}_2 = \begin{bmatrix} .963 \\ .898 \\ .898 \\ 1.021 \\ 1.029 \\ .945 \end{bmatrix}, \text{ and}$$

$$\Pi_3 : \boldsymbol{\mu}_3 = \begin{bmatrix} -1.69 \\ -1.69 \\ -1.69 \\ -1.69 \\ -1.16 \\ -1.16 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2.5 & 1.2 & 1.2 & 1.2 & 1.2 & 1.2 \\ 1.2 & 2.5 & 1.2 & 1.2 & 1.2 & 1.2 \\ 1.2 & 1.2 & 2.5 & 1.2 & 1.2 & 1.2 \\ 1.2 & 1.2 & 1.2 & 2.5 & 1.2 & 1.2 \\ 1.2 & 1.2 & 1.2 & 1.2 & 1.4 & 1.2 \\ 1.2 & 1.2 & 1.2 & 1.2 & 1.2 & 1.2 \end{bmatrix}, \boldsymbol{\gamma}_3 = \begin{bmatrix} 1.063 \\ .998 \\ .998 \\ 1.121 \\ 1.129 \\ 1.045 \end{bmatrix}$$

with $\delta_{0i} = 0$, $i = 1, 2, 3$. As an aside we remark that $\boldsymbol{\gamma}_1 \in \mathcal{N}(\mathbf{C})$. Using Theorem 2.5 to formulate a dimension reduction matrix, we obtain

$$\mathbf{M} = \begin{bmatrix} .385 & -.4 & .3 & .3 & .3 & .3 & .3 & .3 & .5 & .5 & .5 & .5 & .5 & .5 & .2 & .3 \\ 1.154 & -.4 & .3 & .3 & .3 & .3 & .3 & .3 & .5 & .5 & .5 & .5 & .5 & .5 & .2 & .3 \\ 1.154 & -.4 & .3 & .3 & .3 & .3 & .3 & .3 & .5 & .5 & .5 & .5 & .5 & .5 & .2 & .3 \\ -.385 & -.4 & .3 & .3 & .3 & .3 & .3 & .3 & .5 & .5 & .5 & .5 & .5 & .5 & .2 & .3 \\ -.618 & -.4 & .3 & .3 & .3 & .3 & .3 & .3 & .5 & .5 & .5 & .5 & .5 & .5 & .2 & .3 \\ -.618 & -.4 & .3 & .3 & .3 & .3 & .3 & .3 & .5 & .5 & .5 & .5 & .5 & .5 & .2 & .3 \end{bmatrix},$$

where $rank\,(\mathbf{M}) = 2$. Therefore, by Theorem 3.4, the original six-dimensional singular skew-normal density functions can be transformed into two-dimensional density functions without increasing the probability of misclassification under the Bayes classification assignments. Because $\mathcal{C}\,(\mathbf{F}^{+}) = \mathcal{C}\,(\mathbf{F}')$, an optimal two-dimensional representation space is $\mathcal{C}\,(\mathbf{F})$, where

$$\mathbf{F}' = \begin{bmatrix} .415 & .447 & .447 & .384 & .375 & .375 \\ .078 & .486 & .486 & -.330 & -.454 & -.454 \end{bmatrix}.$$

For Π_1, we now have the reduced parameters

$$\boldsymbol{\mu}_{1R} = \begin{bmatrix} 2.44 \\ -.19 \end{bmatrix}, \boldsymbol{\Sigma}_{1R} = \begin{bmatrix} 5.11 & .12 \\ .12 & .79 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{1R} = \begin{bmatrix} 1.84 \\ -.25 \end{bmatrix}.$$

For Π_2, the reduced parameters are

$$\boldsymbol{\mu}_{2R} = \begin{bmatrix} 7.46 \\ 1.19 \end{bmatrix}, \boldsymbol{\Sigma}_{2R} = \begin{bmatrix} 6.90 & -.02 \\ -.02 & .80 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{2R} = \begin{bmatrix} 2.33 \\ -.28 \end{bmatrix}.$$

The reduced parameters for Π_3 are

$$\boldsymbol{\mu}_{3R} = \begin{bmatrix} -3.73 \\ -.16 \end{bmatrix}, \boldsymbol{\Sigma}_{3R} = \begin{bmatrix} 8.09 & -.11 \\ -.11 & .815 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{3R} = \begin{bmatrix} 2.58 \\ -.30 \end{bmatrix}.$$

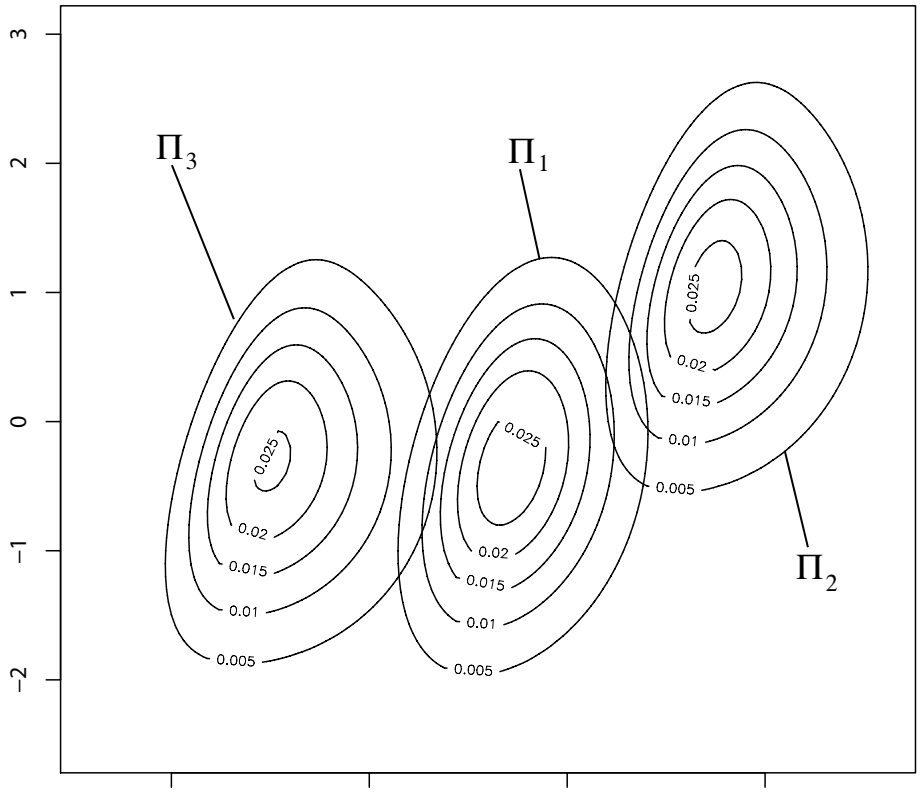

Figure 3.3: The optimal two-dimensional representation for the three six-dimensional singular skew-normal populations given in Example 3.2.

Immediately above, we portray a two-dimensional representation of the original seven-dimensional SSN populations. By using the SVD, we can obtain a one-dimensional representation of both populations; however, we lose some discriminatory information by applying this procedure. For the first population, we obtain parameters

$$\Pi_1 : \mu_{1R} = 2.44, \sigma_{1R} = 5.11, \text{ and } \gamma_{1R} = 1.84.$$

Also, for the second population, we obtain parameters

$$\Pi_2 : \mu_{2R} = 7.46, \sigma_{2R} = 6.90, \text{ and } \gamma_{2R} = 2.33.$$

Finally, for the third population, we obtain parameters

$$\Pi_3 : \mu_{3R} = -3.73, \sigma_{3R} = 8.09, \text{ and } \gamma_{3R} = 2.58.$$

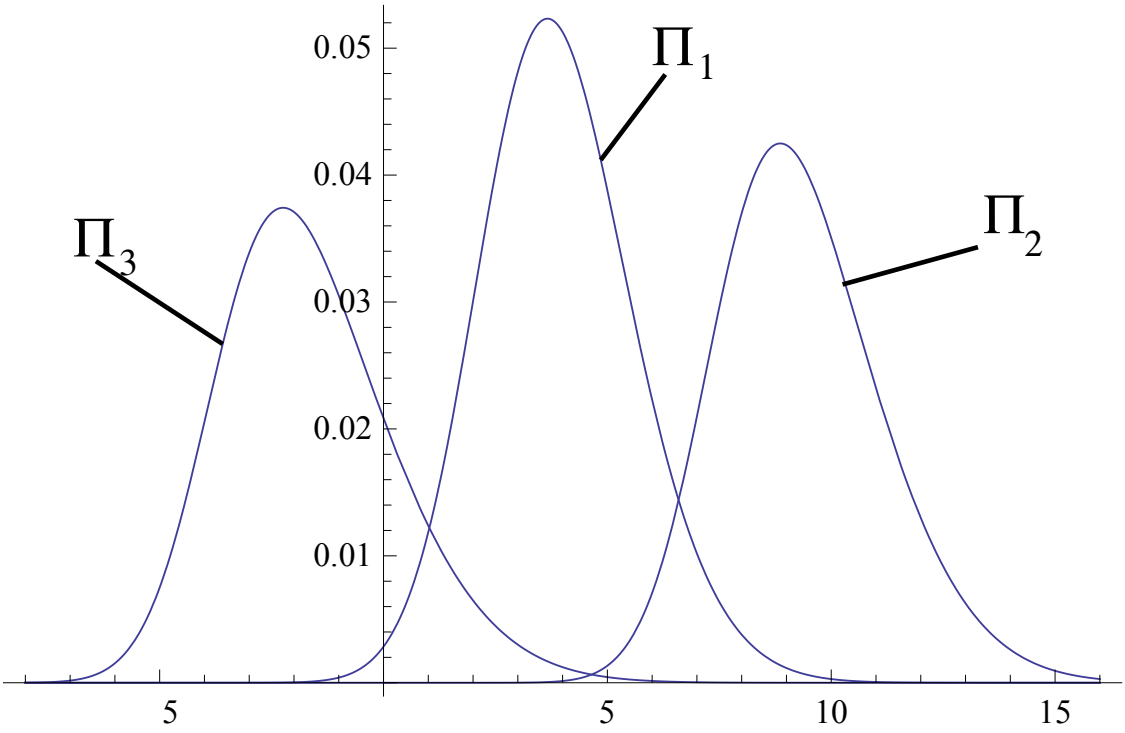

Figure 3.4: A one-dimensional approximate representation for the three six-dimensional singular skew-normal populations in Example 3.2.

3.4 Conclusion

Theorems 3.3 and 3.4 extend the results given in Theorems 2.2 through 2.5 from the nonsingular to the singular case for the dispersion parameters $\mathbf{\Sigma}_i$, $i = 1, 2, ..., m$. If the conditions of the theorem hold, linear dimension reduction from p dimensions to q dimensions, where $1 \leq q < p$ is possible without diminishing the $BPMC$ even though the assumed dispersion parameters $\mathbf{\Sigma}_i$, $i = 1, 2, ..., m$ are singular and the necessary restrictions on the other *MSSN* parameters hold.

CHAPTER FOUR

A Comparison of Two Methods for Linear Discriminant Analysis Using Monotone Missing Training Data

4.1 Introduction

In this chapter, we consider the problem of classifying an unlabeled observation $\mathbf{x} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ into one of two distinct populations, $\Pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. The most well-known statistical procedure for classifying a complete observation vector, assuming the unlabeled observation comes from one of the two multivariate populations with equal covariance matrices, is Fisher's linear discriminant function derived by R. A. Fisher (1936). For the case of unknown parameters, Anderson's linear discriminant function ($ALDF$) (Anderson (1951)) is

$$W = \left[\mathbf{x} - \frac{1}{2}\left(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2\right)\right]' \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right), \tag{4.1}$$

where $\bar{\mathbf{x}}_i$ and $\mathbf{S}$ are estimators of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}$, respectively, such that

$$\bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} \mathbf{x}_{ij}$$

for $i = 1, 2$, and

$$\mathbf{S}_x = \left(\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\right)\left(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\right)'\right)/\nu_x, \qquad \nu_x = n_1 + n_2 - 2.$$

Using (4.1), we classify the unlabeled observation vector $\mathbf{x}$ into class Π_1 if

$$\left(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1\right)' \mathbf{S}^{-1}\mathbf{x} \leq \frac{1}{2}\left(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1\right)' \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_1\right)$$

and classify $\mathbf{x}$ into class Π_2 if (4.2)

$$\left(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1\right)' \mathbf{S}^{-1}\mathbf{x} > \frac{1}{2}\left(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}}_1\right)' \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_2 + \bar{\mathbf{x}}_1\right).$$

The conditional probability of misclassification for classifying an observation from Π_i into Π_{3-i} by W is

$$CER_{i,3-i}\left(ALDF\right) = \Phi\left(\frac{\left(-1\right)^{3-i}\cdot\frac{1}{2}\left(\bar{\mathbf{x}}_1+\bar{\mathbf{x}}_2\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)+\left(-1\right)^{i}\boldsymbol{\mu}_i'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)}{\sqrt{\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)'\mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2\right)}}\right) \tag{4.3}$$

for $i = 1, 2$, where $\Phi\left(\cdot\right)$ denotes the cumulative distribution function of the univariate standard normal distribution. Thus, assuming equal *a priori* class membership probabilities, the conditional error rate for $ALDF$ is

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

The expected error rate (EER) for $ALDF$ given in (4.2) is

$$EER\left(ALDF\right) = \frac{1}{2}\left[P\left(W<0|\mathbf{x}\in\Pi_1\right) + P\left(W\geq 0|\mathbf{x}\in\Pi_2\right)\right]. \tag{4.4}$$

An expression for $EER\left(ALDF\right)$ has been developed by John (1961) but is in the form of an infinite sum.

Jackson (1968) has considered the problem of missing values in a discriminant function analysis where the numbers of both variables and individuals are very large. Estimation of missing values using mean substitution and estimation by an iterative regression technique are assayed and the results compared. She concludes that the far simpler method of mean substitution and the iterative regression technique give similar results. Additionally, Chan and Dunn (1972) have examined the probability of correct classification under the most popular methods of handling data values that are missing at random. The EER is used as a criterion to weigh the relative quality of supervised classification methods. Moreover, Chan, Gilman, and Dunn (1976) have handled missing observations in discrimination for a variety of population covariance matrices using a regression technique and a modified technique contingent on the first principal component. Furthermore, Titterington and Jiang (1983) have applied recursive methods for handling incomplete data and have verified asymptotic properties for the recursive methods.

The missing data pattern we address in this paper is monotone missing data. Monotone missing data occurs for data vector $\mathbf{x}_j$ in the case that if $\mathbf{x}_{ji}$ is missing, then $\mathbf{x}_{jk}$ is missing for all $k > i$. Chung and Han (2000) have performed a Monte Carlo simulation in which their classifier performs better than an MLE classifier formulated by Bohannon and Smith (1975) in terms of the expected error rate in all situations despite the fact that the MLE classifier incorporates the correlation among the variables with missing and non-missing observations. We demonstrate that the MLE classifier can considerably outperform the *C-H* classifier in terms of their respective EERs for certain covariance matrix configurations, especially covariance matrices with moderate to high correlation among the features with no missing data and the features with monotone missing data.

We have organized this chapter as follows. In Section 2, we describe the *C-H* classifier, a linear combination discriminant analysis procedure from Chung and Han (2000) when the training data from both classes contain identical monotone missing data patterns. Also, we describe the MLE-based linear discriminant procedure from Batsidis, Zografos, and Loukas (2006) when the training sets from both classes contain identical monotone missing data patterns. We also derive the MLEs for the two means and common covariance matrix of two p-dimensional normal distributions with identical monotone missing data patterns. In Section 3, we perform two Monte Carlo simulations to examine the differences in the estimated EERs of the *C-H* and MLE linear 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 MLE linear classifiers using parametric bootstrap EER estimators on two actual data sets. Finally, we summarize the results of the two simulation and two real-data comparisons of the *C-H* and MLE classifiers in Section 5.

4.2 Two Competing Classifiers for Monotone Missing Training Data

4.2.1 The C-H Classifier for Monotone Missing Data

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

$$\begin{bmatrix} \mathbf{Y}_{i1} & \mathbf{Y}_{i2} \\ \mathbf{Z}_i & \cdot \end{bmatrix}, \tag{4.5}$$

where

$$\mathbf{U}_i = \left[\mathbf{Y}'_{i1}\ ,\ \mathbf{Z}'_i\right]' \in \mathbb{R}_{k\times n_i} \tag{4.6}$$

denotes the n_i complete-observation matrix, and $\mathbf{Y}_{i2} \in \mathbb{R}_{k\times(N_i-n_i)}$ is the partial observation matrix whose first k measurements are non-missing, where $N_i > n_i$, for $i = 1, 2$.

Chung and Han (2000) have derived a linear combination of a discriminant function composed from complete data and a discriminant function determined from incomplete data in the form of monotone missing data. We denote the complete data by $\mathbf{u}_{ij} = \left[\mathbf{y}'_{i1j}, \mathbf{z}'_{ij}\right]'$, where $\mathbf{y}_{i1j} \in \mathbb{R}_{k\times 1}$ and $\mathbf{z}_{ij} \in \mathbb{R}_{(p-k)\times 1}$ such that

$$\mathbf{u}_{ij} \sim N_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right) \equiv N_p\left(\begin{bmatrix} \boldsymbol{\mu}_{Y_{i1}} \\ \boldsymbol{\mu}_{Z_i} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right), \tag{4.7}$$

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 $i = 1, 2;\ j = 1, 2, ..., n_i$. 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 $\boldsymbol{\Sigma}_{yy} \in \mathbb{R}^{>}_{k\times k}$. Anderson's discriminant function of the subset of complete data $\mathbf{U}_i$, $i = 1, 2$, given in (4.6), is

$$W_u \equiv \left(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2\right)' \mathbf{S}_u^{-1}\left[\mathbf{u} - \frac{1}{2}\left(\bar{\mathbf{u}}_1 + \bar{\mathbf{u}}_2\right)\right], \tag{4.8}$$

such that

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

where $\bar{\mathbf{u}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{u}_{ij}$ is the complete-data sample mean and $\mathbf{S}_u$ is the complete-data sample covariance matrix. Anderson's discriminant function for the data

$$[\mathbf{Y}_{i1} : \mathbf{Y}_{i2}], \tag{4.9}$$

$i = 1, 2$, the k-dimensional features with $N_1 + N_2$ training observations, is

$$W_y \equiv (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_y^{-1} \left[\mathbf{y} - \frac{1}{2}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2)\right] \tag{4.10}$$

with

$$\bar{\mathbf{y}}_i = \frac{1}{N_i}\left[n_i \bar{\mathbf{y}}_{i1} + (N_i - n_i)\,\bar{\mathbf{y}}_{i2}\right], \tag{4.11}$$

where

$$\bar{\mathbf{y}}_{i1} = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{y}_{i1j} \tag{4.12}$$

denotes the sample mean for the first n_i observations and the first k features in (4.5),

$$\bar{\mathbf{y}}_{i2} = \frac{1}{N_i - n_i}\sum_{j=n_i+1}^{N_i}\mathbf{y}_{i2j} \tag{4.13}$$

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

$$\mathbf{S}_y = \left(\sum_{i=1}^{2}\sum_{k=1}^{2}\sum_{j=1}^{N_i}(\mathbf{y}_{ikj} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ikj} - \bar{\mathbf{y}}_i)'\right)/\nu_y, \quad \nu_y = N_1 + N_2 - 2,$$

is the pooled sample covariance matrix for the incomplete training data (4.9), where $i = 1, 2$.

Chung and Han (2000) have proposed the linear combination classification statistic

$$W_c \equiv cW_u + (1 - c)\,W_y, \tag{4.14}$$

where $c \in [0, 1]$. We classify the unlabeled observation vector $\mathbf{x} \in \mathbb{R}_{p\times 1}$ into Π_1 if

$$W_c \geq 0 \tag{4.15}$$

and into Π_2, otherwise. The conditional error rate (CER) for classifying an unlabeled vector $\mathbf{x}$ from Π_1 into Π_2 using W_c is

$$\begin{aligned} CER_{12}(W_c) &= P(W_c < 0 | \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, \mathbf{S}_u, \bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \mathbf{S}_y; \mathbf{u}, \mathbf{y} \in \Pi_1) \\ &= \Phi\left(\frac{-\mathbf{h}'\boldsymbol{\mu}_1 - f}{\sqrt{\mathbf{h}'\boldsymbol{\Sigma}\mathbf{h}}}\right), \end{aligned} \tag{4.16}$$

where

$$f \equiv cb + (1-c)\, e \tag{4.17}$$

with

$$\begin{aligned} b &\equiv -\frac{1}{2}(\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)' \mathbf{S}_u^{-1}(\bar{\mathbf{u}}_1 + \bar{\mathbf{u}}_2), \\ e &\equiv -\frac{1}{2}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_y^{-1}(\bar{\mathbf{y}}_1 + \bar{\mathbf{y}}_2), \end{aligned}$$

and

$$\mathbf{h} \equiv \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}, \tag{4.18}$$

where $\mathbf{a} \in \mathbb{R}_{k\times 1}$ and $\mathbf{b} \in \mathbb{R}_{(p-k)\times 1}$ such that

$$\begin{aligned} \mathbf{a} &\equiv c\mathbf{a}_1 + (1-c)\,\mathbf{d}, \\ \mathbf{b} &\equiv c\mathbf{a}_2, \\ \mathbf{d}' &\equiv (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}_y^{-1}, \\ \mathbf{a}' &\equiv (\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)' \mathbf{S}_u^{-1} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{bmatrix}, \end{aligned}$$

where $\mathbf{a}_1 \in \mathbb{R}_{k\times 1}$ and $\mathbf{a}_2 \in \mathbb{R}_{(p-k)\times 1}$. Similarly,

$$\begin{aligned} CER_{21}(W_c) &= P(W_c \geq 0 | \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2, \mathbf{S}_u, \bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \mathbf{S}_y; \mathbf{u}, \mathbf{y} \in \Pi_2) \\ &= \Phi\left(\frac{\mathbf{h}'\boldsymbol{\mu}_2 + f}{\sqrt{\mathbf{h}'\boldsymbol{\Sigma}\mathbf{h}}}\right). \end{aligned} \tag{4.19}$$

Thus, assuming equal prior probability, the CER for the C-H classifier (4.14) is defined to be

$$CER(W_c) = \frac{1}{2}\left[CER_{12}(W_c) + CER_{21}(W_c)\right]. \tag{4.20}$$

If

$$\tilde{\boldsymbol{\theta}} \equiv [\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2, \mathbf{S}_y, \mathbf{S}_u, \bar{\mathbf{u}}_1, \bar{\mathbf{u}}_2]',$$

then, for the classifier (4.15), the EER of misclassifying an unlabeled observation vector $\mathbf{x}$ from Π_1 into Π_2 is

$$EER\,(W_c)_{12} = E_{\tilde{\theta}}\left[\Phi\left(\frac{-\mathbf{h}'\boldsymbol{\mu}_1 - f}{\sqrt{\mathbf{h}'\boldsymbol{\Sigma}\mathbf{h}}}\right)\right],$$

and, similarly, the EER of misclassifying $\mathbf{x}$ from Π_2 into Π_1 is

$$EER\,(W_c)_{21} = E_{\tilde{\theta}}\left[\Phi\left(\frac{\mathbf{h}'\boldsymbol{\mu}_2 + f}{\sqrt{\mathbf{h}'\boldsymbol{\Sigma}\mathbf{h}}}\right)\right].$$

Thus, assuming equal prior probabilities, the EER for (4.15) is

$$EER\,(W_c) = \frac{1}{2}\left[EER_{12}\,(W_c) + EER_{21}\,(W_c)\right]. \tag{4.21}$$

In choosing c in (4.14), Chung and Han (2000) 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 deals with small CERs, the sample Mahalanobis distance D_k^2, $k = u, y$ for either the difference between the $\mathbf{u}_i$'s or $\mathbf{z}_i$'s for $i = 1, 2$, will be large or the training sample sizes will be large. While n_i and D_u^2 determine the performance of W_u, the quantities N_i and D_y^2 dictate the performance of W_y. Hence, Chung and Han (2000) have chosen c in relation to the training-sample sizes and the Mahalanobis distances for the complete and incomplete training-data sets. The implication for times when D_u^2 is larger than D_y^2 is that the information in the data-matrix component $\mathbf{Z}_i$, $i = 1, 2$, in (4.5) contributes largely to the discriminatory information. Hence, Chung and Han (2000) use

$$c^* = \frac{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} D_u^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} D_u^2 + \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^{-1} D_y^2}, \tag{4.22}$$

where

$$D_y^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)'\,\mathbf{S}_y^{-1}\,(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \tag{4.23}$$

and

$$D_u^2 = (\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2)' \, \mathbf{S}_u^{-1} \, (\bar{\mathbf{u}}_1 - \bar{\mathbf{u}}_2) \,, \tag{4.24}$$

to determine the linear combination classifier (4.14).

4.2.2 The Maximum Likelihood Classifier for Monotone Missing Training Data

Anderson (1957) has examined maximum likelihood estimators (MLEs) for the parameter of a multivariate normal distribution for special patterns of missing observations in the training samples. Also, Hocking and Smith (1968) have derived an MLE method for estimating parameters in a multivariate normal distribution with monotone missing data. Once computed, the MLEs are substituted for the parameters in the optimal Bayes classifier. However, the estimator of $\mathbf{\Sigma}$ in the Hocking and Smith (1968) MLE classifier is a pooled estimator of the two individual MLE estimators of $\mathbf{\Sigma}$.

Below, we state and derive the MLEs for two multivariate normal distributions having unequal means and a common covariance matrix using identical monotone missing-data patterns in both training samples. We give a matrix-calculus-based proof of the following theorem in Appendix C.

Theorem 4.1. Let Π_i be modeled with the multivariate normal densities $N_p(\boldsymbol{\mu}_i, \mathbf{\Sigma})$ for $i = 1, 2$, with

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_{i1} \\ \boldsymbol{\mu}_{i2} \end{bmatrix} \tag{4.25}$$

and

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{21} \\ \mathbf{\Sigma}_{12} & \mathbf{\Sigma}_{22} \end{bmatrix}, \tag{4.26}$$

and let

$$\mathbf{A}_{11,N_i,i} = \sum_{j=1}^{N_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)' \,, \tag{4.27}$$

$$\mathbf{A}_{11,n_i,i} = \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)' \,, \tag{4.28}$$

$$\mathbf{A}_{12,n_i,i} = \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)', \tag{4.29}$$

and

$$\mathbf{A}_{22,n_i,i} = \sum_{j=1}^{n_i} (\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)(\mathbf{z}_{ij} - \bar{\mathbf{z}}_i)', \tag{4.30}$$

where $\mathbf{y}_{ij} \in [\mathbf{Y}_{i1} : \mathbf{Y}_{i2}]$, and $\mathbf{z}_{ij} \in \mathbf{Z}_i$ with $\mathbf{Y}_{i1}$, $\mathbf{Y}_{i2}$, and $\mathbf{Z}_i$ given in (4.5). Then, on the basis of two-step monotone training samples from populations $\Pi_i : N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, the MLEs of (4.25) and (4.26) are

$$\hat{\boldsymbol{\mu}}_i = \begin{bmatrix} \hat{\boldsymbol{\mu}}_{i1} \\ \hat{\boldsymbol{\mu}}_{i2} \end{bmatrix} \text{ and } \widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \widehat{\boldsymbol{\Sigma}}_{11} & \widehat{\boldsymbol{\Sigma}}_{12} \\ \widehat{\boldsymbol{\Sigma}}_{21} & \widehat{\boldsymbol{\Sigma}}_{22} \end{bmatrix}, \tag{4.31}$$

respectively, where

$$\widehat{\boldsymbol{\Sigma}}_{11} = \frac{\sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}}{\sum_{i=1}^{2} N_i}, \tag{4.32}$$

$$\widehat{\boldsymbol{\Sigma}}_{12} = \frac{1}{\left(\sum_{i=1}^{2} N_i\right)} \left[\sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right], \tag{4.33}$$

and

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{22} =& \frac{1}{\sum_{i=1}^{2} n_i} \sum_{i=1}^{2} \mathbf{A}_{22\cdot 1,n_i,i} + \frac{1}{\sum_{i=1}^{2} N_i} \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}\right] \times \\ & \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right], \end{aligned} \tag{4.34}$$

and $\hat{\boldsymbol{\mu}}_{i1} = \bar{\mathbf{y}}_i$ with $\bar{\mathbf{y}}_i$ defined in (4.11),

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

where

$$\bar{\mathbf{z}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{z}_{ij},$$

and $\widehat{\boldsymbol{\Sigma}}_{12}$, $\widehat{\boldsymbol{\Sigma}}_{22}$, $\bar{\mathbf{y}}_{i1}$, and $\bar{\mathbf{y}}_{i2}$ are defined in (4.33), (4.34), (4.11), and (4.12), respectively, for $i = 1, 2$.

Proof: The proof is given in Appendix C.

The *MLE* classification statistic is

$$W_{MLE} \equiv (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)' \, \widehat{\boldsymbol{\Sigma}}^{-1} \left[\mathbf{x} - \frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1) \right], \tag{4.35}$$

where $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$ are the *MLE*s defined in (4.31), and $\mathbf{x} \in \mathbb{R}_{p\times 1}$ is an unlabeled observation vector belonging to either Π_1 or Π_2. We classify the unlabeled observation vector $\mathbf{x} \in \mathbb{R}_{p\times 1}$ into class Π_1 if

$$W_{MLE} \leq 0 \tag{4.36}$$

and into Π_2, otherwise. Hence, from (4.31) and (4.2) and conditioning on $\hat{\boldsymbol{\mu}}_{ij}$, $i = 1, 2$, and $\widehat{\boldsymbol{\Sigma}}$, we have

$$CER_{12} \left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}} \right) \equiv P \left[W_{MLE} > 0 | \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}, \mathbf{x} \in \Pi_1 \right], \tag{4.37}$$

where $\mathbf{x}$ is a complete unlabeled observation. Given $\mathbf{x} \in \Pi_1$ and $\hat{\delta} \equiv \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$, we have

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

which implies

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

where

$$w_1 = \left[\hat{\delta}' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\delta} \right]^{-1/2} \left\{ \hat{\delta}' \widehat{\boldsymbol{\Sigma}}^{-1} \left(\frac{1}{2} (\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1) - \boldsymbol{\mu}_1 \right) \right\}.$$

Similarly, given $\mathbf{x} \in \Pi_2$,

$$\hat{\delta}' \widehat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \sim N \left(0, (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1)' \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \widehat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) \right),$$

which implies

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

where

$$w_2 = \left[\hat{\delta}'\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}^{-1}\hat{\delta}\right]^{-1/2}\left\{\hat{\delta}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_2+\hat{\boldsymbol{\mu}}_1\right)-\boldsymbol{\mu}_2\right)\right\}.$$

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

$$CER\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right) = \frac{1}{2}\left[CER_{12}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right) + CER_{21}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right)\right]. \tag{4.38}$$

In addition, the expected error rate for $\mathbf{x} \in \Pi_1$ is

$$EER_{12}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right) = 1 - E_{\tilde{\boldsymbol{\theta}}}\left(\Phi\left(w_1\right)\right),$$

and the expected error rate for $\mathbf{x} \in \Pi_2$ is

$$EER_{21}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right) = E_{\tilde{\boldsymbol{\theta}}}\left(\Phi\left(w_2\right)\right).$$

Hence, the overall expected error rate is

$$EER\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right) = \frac{1}{2}\left[EER_{12}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right) + EER_{21}\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\boldsymbol{\Sigma}}\right)\right].$$

4.3 Two Monte Carlo Simulations

In this section, we present a description and the results of two Monte Carlo simulations we have performed to analyze the difference in the estimated expected error rates $\left(\widehat{EER}\text{s}\right)$ of the *MLE* and *C-H* classifiers for two multivariate normal configurations, where $\Pi_i : N_p\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right)$, $i = 1, 2$, using various training-sample sizes and two 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

$$\boldsymbol{\Sigma} \equiv \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

is the intraclass covariance matrix where ρ denotes the population common covariance matrix among the features in the intraclass covariance matrix.

The two simulations have been performed in SAS 9.2 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 $C\text{-}H$ classifiers have been computed, and their CERs have been calculated for each training sample set. Then, the differences between the CERs for the classifiers have been averaged over the 10,000 CER differences for the two classifiers for each parameter configuration involving N_i, p, r, $\mathbf{\Sigma}$, $\boldsymbol{\mu}_i$ and the percent of observations with missing data ($POMD$) for the r features with monotone missing data, where $i = 1, 2$. Thus, the estimated expected error rate difference for the $C\text{-}H$ and MLE classifiers is

$$\widehat{EERD} = \widehat{EER}_{C\text{-}H} - \widehat{EER}_{MLE}, \tag{4.39}$$

where

$$\widehat{EER}_{C\text{-}H} = \frac{1}{k}\sum_{j=1}^{k} CER_j\left(W_c\right)$$

is the estimated expected error rate for the $C\text{-}H$ classifier and

$$\widehat{EER}_{MLE} = \frac{1}{k}\sum_{j=1}^{k} CER_j\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\mathbf{\Sigma}}\right)$$

is the estimated expected error rate for the MLE classifier. Also, $CER\left(W_c\right)$ is defined in (4.20), $CER\left(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \widehat{\mathbf{\Sigma}}\right)$ is given in (4.38), k is the total number of simulated training-data sets, and j denotes the j^{th} simulated training-data set where $j \in \{1, 2, ..., k\}$. We display the results of our two Monte Carlo simulations by graphing $\widehat{EERD}$ against various values of ρ for fixed values of p, r, N_i, d, and $POMD$ for the r features with monotone missing data.

The relationship between p and r has been fixed at $r = .2p$ and $r = .8p$. We have chosen these specific values of p and r to evaluate the $\widehat{EERD}$ when small and large proportions of variables with missing data exist relative to p, the dimension

of the training sample. Hence, for $p = 10$, we have used $r = 2, 8$; for $p = 20$, we have used $r = 4, 16$; and for $p = 40$, we have used $r = 8, 32$. Additionally, we have used sample sizes of $N_i = 20, 50, 100$ when $p = 10$; $N_i = 25, 50, 100$ when $p = 20$; and $N_i = 50, 100, 200$ when $p = 40$, $i = 1, 2$. The total sample sizes N_i represent small, medium, and large sample sizes relative to the specified dimension p. Lastly, we mention that for this simulation, we have chosen $\boldsymbol{\mu}_1 \in \mathbb{R}_{p\times 1}$ such that

$$\boldsymbol{\mu}_1 = [0, 0, ..., 0]' \tag{4.40}$$

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

$$\boldsymbol{\mu}_2 = [d_j, 0, 0, ..., 0, d_j, 0, ...0]', \tag{4.41}$$

with $d_1 = .5$ and $d_2 = 3$ to assess $\widehat{EERD}$ for both small and large between-classes separation, where 0.20 of the values in $\boldsymbol{\mu}_2$ are non-zero. As in Chung and Han (2000), we have chosen $N_i > p$ to avoid singularity of the covariance matrices. Furthermore, we have contrasted the two classifiers (4.15) and (4.36) with the $POMD = .5, .8$ for the r variables with monotone missing data. The comparison criterion $\widehat{EERD}$ is plotted against ρ for various combinations of p, r, d, N_i, and $POMD$ in Figures 4.1 - 4.3. For each combination of the parameters mentioned immediately above, we graph $\widehat{EERD}$ versus the intraclass covariance values $\rho = .1, .3, .5, .7, .9$. These values for $\boldsymbol{\mu}_i$, $i = 1, 2$, defined in (4.40) and (4.41), are similar to the population means used in the simulation in Chung and Han (2000).

Figures 4.1 - 4.3 illustrate that the $\widehat{EERD}$ is consistently positive for the values of p, r, N_i, ρ, and $POMD$ examined here. Thus, the $\widehat{EER}_{MLE} < \widehat{EER}_{C\text{-}H}$ for the parameter configurations. Moreover, Figures 4.1 - 4.3 indicate that the primary parameter values that verify the dominance of the MLE classifier are ρ and d. For all three values of $p = 10, 20$, and 40, the $C\text{-}H$ and MLE classifiers are competitive for $\rho = .1$; however, the $\widehat{EERD} > 0$, indicating a slightly smaller $\widehat{EER}_{MLE}$. More importantly, as ρ approaches 1, the MLE classifier performs increasingly better than

the *C-H* classifier. The most noteworthy increase in the $\widehat{EERD}$ is for $.7 \leq \rho \leq .9$ when $d = .5$, where the $\widehat{EERD}$ increases by about .10. This increase occurs for all specified values of p, r, N_i, and $POMD$, and, thus, ascertains the superiority of the MLE classifier in terms of the $\widehat{EERD}$ for these configurations. Additionally, we note that $\widehat{EERD} = .20$ when $\rho = .9$ for most parameter configurations when $d = .5$. These results imply that the MLE classifier is increasingly superior to the *C-H* classifier when strong correlation exists.

The MLE classifier especially outperforms the *C-H* classifier when $d_1 = .5$ for all $\rho > .1$ considered here, as compared to when $d_2 = 3$. The smaller difference in the $\widehat{EERD}$ when $d_2 = 3$ can be attributed to the fact that for a large Mahalanobis distance such as when $d_2 = 3$ and $\rho = .1$, the EERs for both techniques are small, thus providing a smaller $\widehat{EERD}$. When little class separation exists, such as when $d_1 = .5$, then $\widehat{EER}_{MLE} < \widehat{EER}_{C\text{-}H}$, but only slightly. The MLE classifier is superior to the *C-H* classifier when $\rho \geq .1$ for all the population and sample sizes considered here because the MLE classifier incorporates the correlation in the training data between the variables with no missing data and the variables with missing data to effectively estimate the multivariate normal parameters whereas the *C-H* classifier discards this information.

As Figures 4.1 - 4.3 indicate, the contrasting values of p, r, N_i, and $POMD$ contribute marginally, if at all, to $\widehat{EERD}$. Regardless of the combination of values for each of the previously mentioned parameters considered here, the MLE classifier dominates the *C-H* classifier in terms of $\widehat{EERD}$.

The notation in the second simulation is identical to that of the first simulation. Also, $\boldsymbol{\mu}_1$ is defined as in (4.40) for the second simulation, and we use two different vectors for $\boldsymbol{\mu}_2$. We use

$$\boldsymbol{\mu}_2 = [d_j, d_j, ..., 0, d_j, d_j, ..., 0]', j = 1, 2,$$

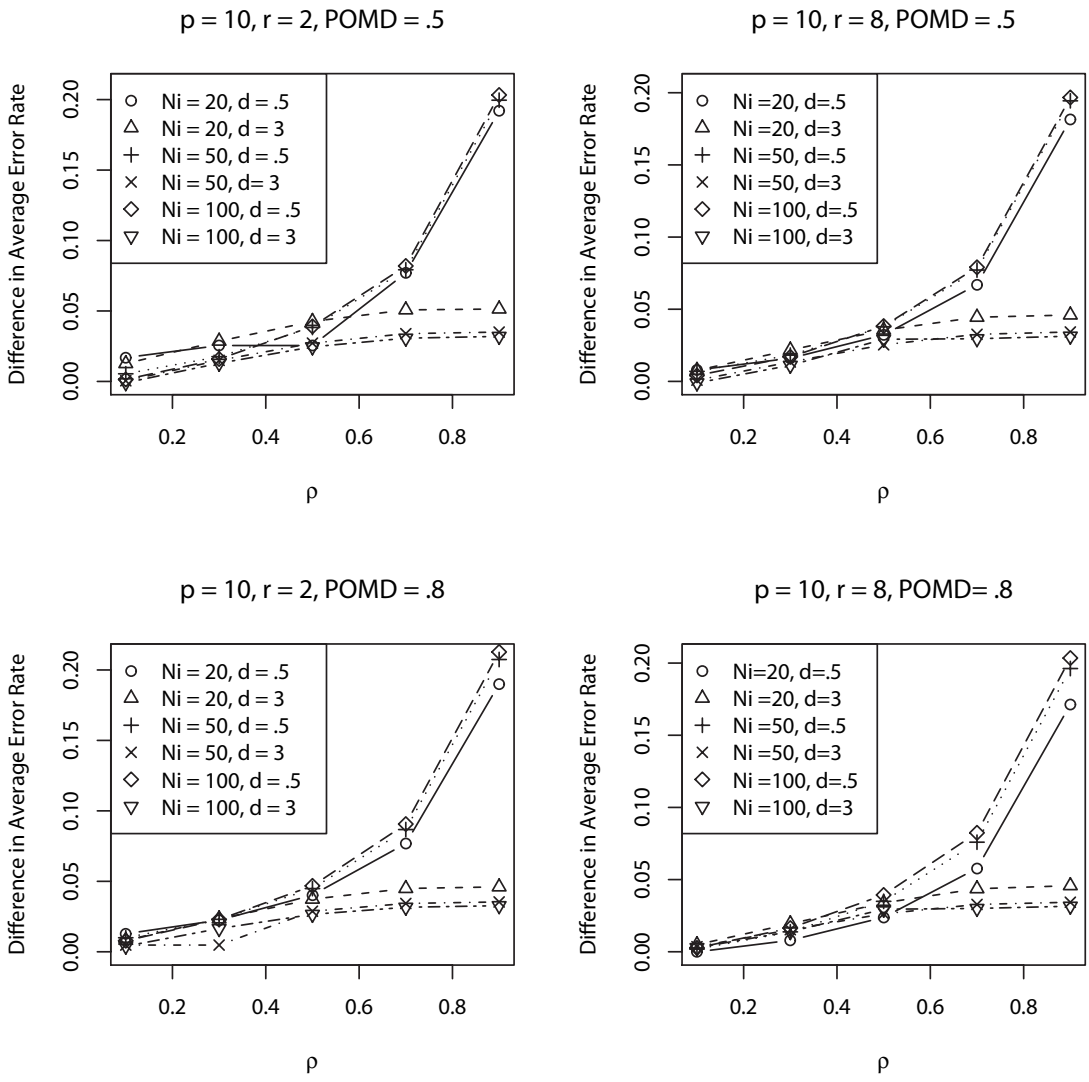

Figure 4.1: Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, $POMD$, and $p = 10$.

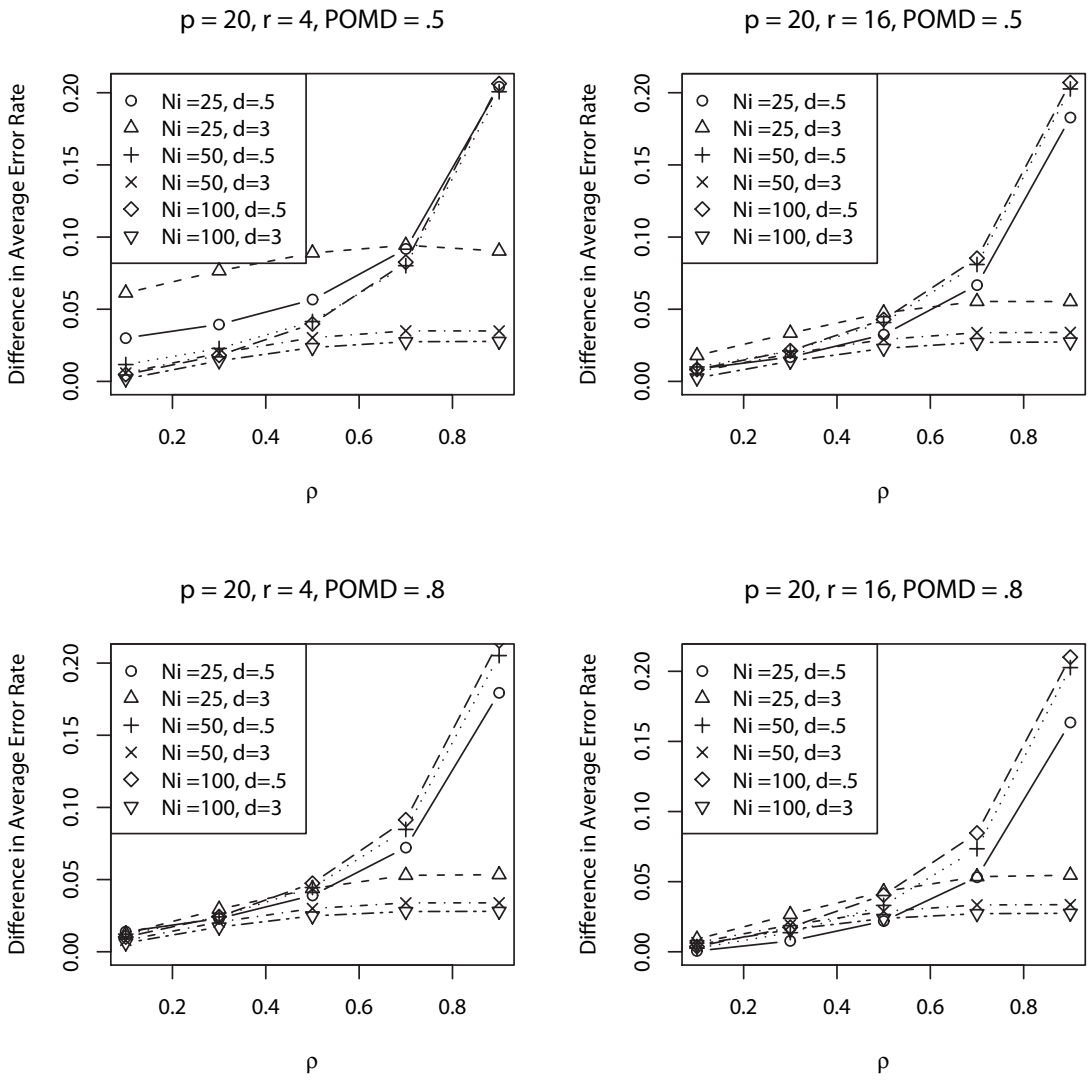

Figure 4.2: Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, $POMD$, and $p = 20$.

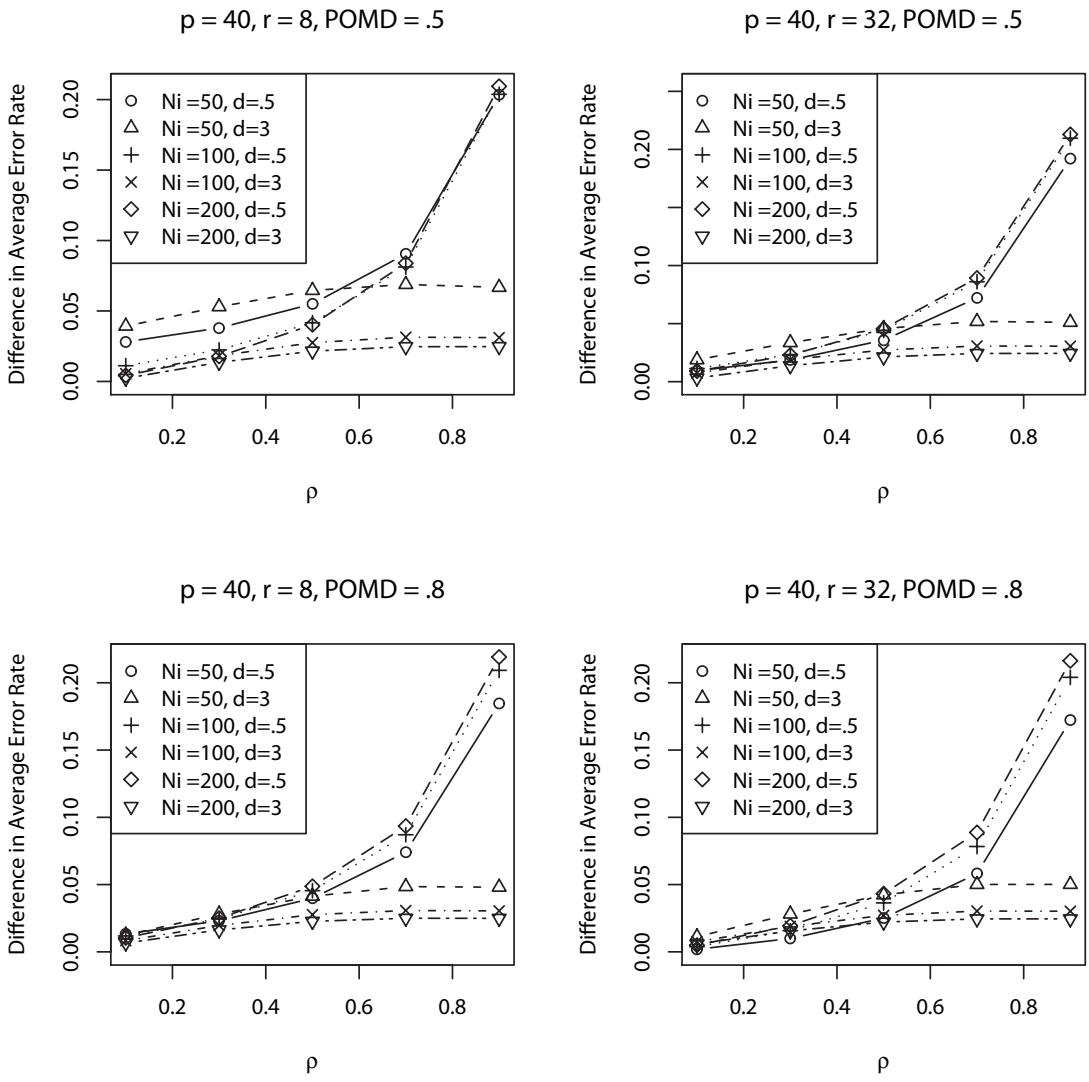

Figure 4.3: Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, $POMD$, and $p = 40$.

such that $d_1 = .125$ and $d_2 = .75$, where 80% of the elements are non-zero. We emphasize that the nonzero elements of $\boldsymbol{\mu}_2$ in the second simulation are different from the nonzero values for $\boldsymbol{\mu}_2$ in the first simulation. This choice of values for d_j, $j = 1, 2$, has been made to maintain that the sum of the elements in $\boldsymbol{\mu}_2$ from the second simulation equals the sum of the elements in $\boldsymbol{\mu}_2$ from (4.41) in the first simulation. The only difference in the sample sizes between the two simulations is the value of N_i for $p = 10$, which has been set to be 25 instead of 20. Consequently, we obtain somewhat different results for the $\widehat{EERD}$ plotted against ρ from the first simulation. One can view the simulation results in Figures 4.4 - 4.6.

In the second simulation, the results suggest that when ρ exceeds .5, the MLE classifier is superior to the *C-H* classifier in terms of the $\widehat{EERD}$. Conclusively, as ρ approaches 1, the MLE classifier becomes increasingly superior in terms of the $\widehat{EERD}$. Here, the level of superiority between the two classifiers also relies on the value of d_j. For $d_2 = .75$ and for $\rho \leq .3$, $\widehat{EERD} < 0$, which implies that the *C-H* classifier performs better on average. However, for $\rho \geq .5$ when $d_2 = .75$, the MLE classifier becomes increasingly superior in terms of the $\widehat{EERD}$ as ρ approaches 1, compared to the case when $d_1 = .125$. The fact that for $\rho \geq .5$, $\widehat{EER}_{MLE} < \widehat{EER}_{C\text{-}H}$ for all parameter configurations suggests that for a moderate to large degree of correlation between the features with no missing data and the features with missing data, the MLE classifier is preferred to the *C-H* classifier. Additionally, we note that $\widehat{EER}_{MLE} \ll \widehat{EER}_{C\text{-}H}$ for certain parameter configurations. In particular, $\widehat{EERD} \approx .20$ when $\rho = .9$ for most parameter configurations with $d = .125$. This result implies that the MLE classifier is much more superior when strong correlation exists between the variables without missing data and those with missing data. We remark that neither the training sample sizes, N_i, $i = 1, 2$, nor the $POMD$ appear to fundamentally affect the relative performance of the two competing classifiers for the parameter configurations considered here.

4.4.1 Bootstrap Expected Error Rate Estimators for the C-H and MLE Classifiers

In this section, we compare a parametric bootstrap estimated EER of the *C-H* and MLE classifiers for two real data sets each having two populations with different population means and equal covariance matrices. First, we define the bootstrap expected error rate estimator for the *C-H* classifier, $\widehat{EER}_{Boot(C\text{-}H)}$. Let $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\widehat{\boldsymbol{\Sigma}}$ be the MLEs of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ defined in Theorem 4.1, respectively. Also, let $\hat{\boldsymbol{\mu}}_1^*$, $\hat{\boldsymbol{\mu}}_2^*$, and $\widehat{\boldsymbol{\Sigma}}^*$ be the bootstrap estimates of $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\widehat{\boldsymbol{\Sigma}}$, respectively, calculated using the bootstrap training-sample data sets

$$\begin{bmatrix} \mathbf{Y}_{i1}^* & \mathbf{Y}_{i2}^* \\ \mathbf{Z}_i^* & \cdot \end{bmatrix}, \tag{4.42}$$

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

$$CER_{12}^*\left(W_c^*\right) = \Phi\left(\frac{-\mathbf{h}^{*\prime}\hat{\boldsymbol{\mu}}_1 - f^*}{\sqrt{\mathbf{h}^{*\prime}\widehat{\boldsymbol{\Sigma}}\mathbf{h}^*}}\right)$$

and

$$CER_{21}^*\left(W_c^*\right) = \Phi\left(\frac{\mathbf{h}^{*\prime}\hat{\boldsymbol{\mu}}_2 + f^*}{\sqrt{\mathbf{h}^{*\prime}\widehat{\boldsymbol{\Sigma}}\mathbf{h}^*}}\right),$$

where W_c^*, $\mathbf{h}^*$, and f^* are similar in definition to W_c, $\mathbf{h}$ and f in (4.15), (4.18), and (4.17), respectively, except that we use the bootstrap multivariate normal data in (4.42). Thus, assuming equal prior probability, the bootstrap CER for the *C-H* classifier is

$$CER^*\left(W_c^*\right) = \frac{1}{2}\left[CER_{12}^*\left(W_c^*\right) + CER_{21}^*\left(W_c^*\right)\right]. \tag{4.43}$$

Then, the estimated bootstrap expected error rate for the *C-H* classifier is

$$\widehat{EER}_{Boot(C\text{-}H)} = \frac{1}{k}\sum_{j=1}^{k} CER_j^*\left(W_c^*\right),$$

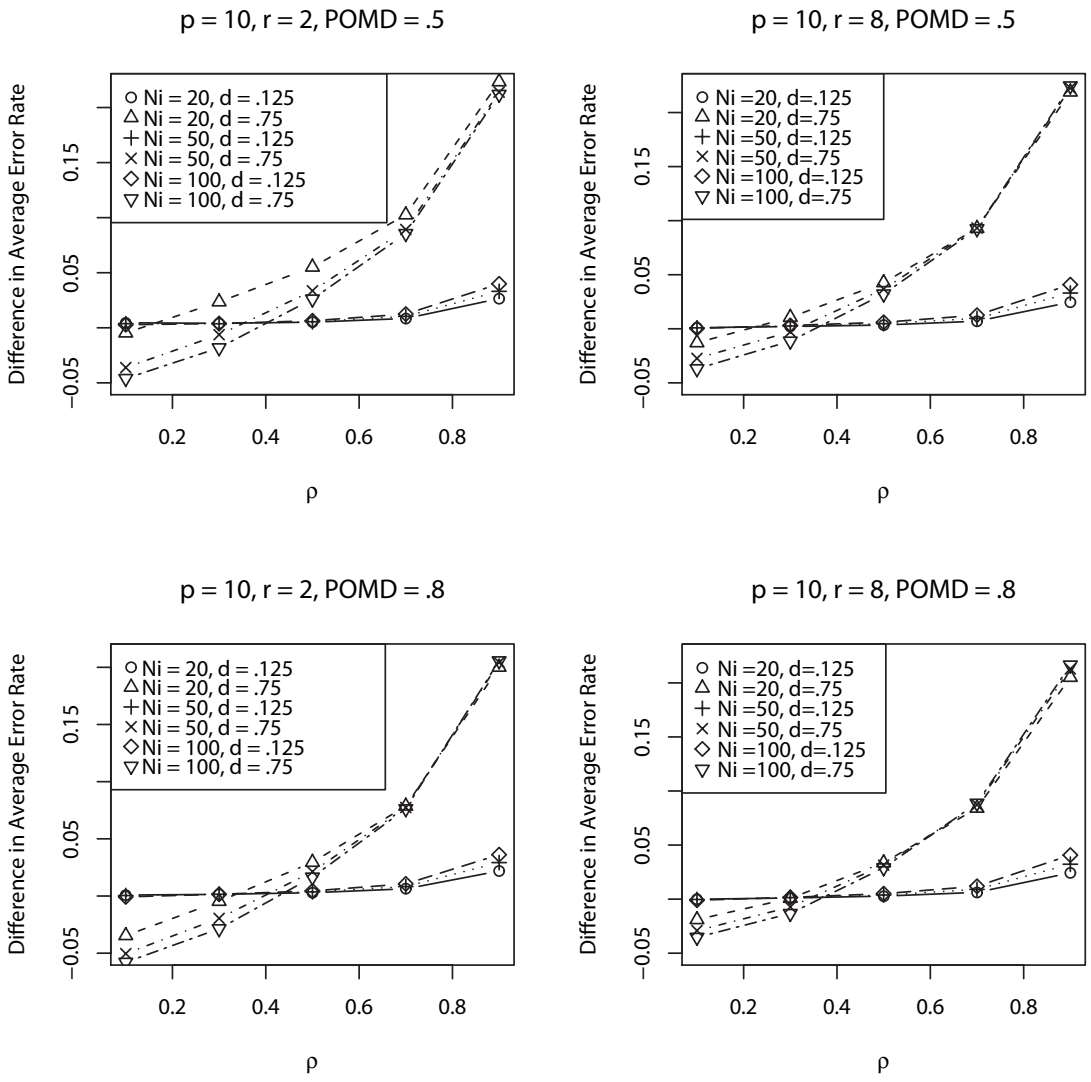

Figure 4.4: Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, $POMD$, and $p = 10$.

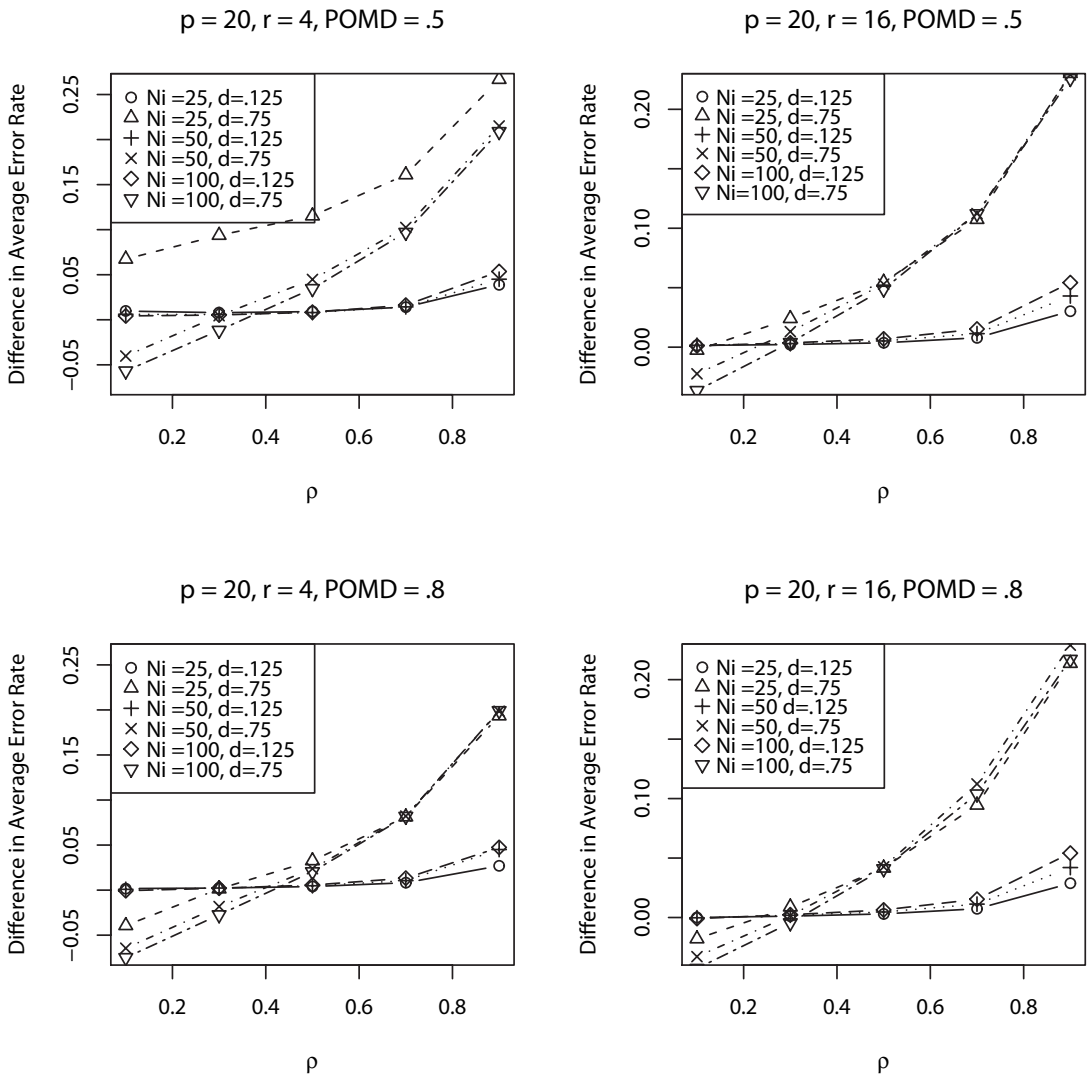

Figure 4.5: Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, $POMD$, and $p = 20$.

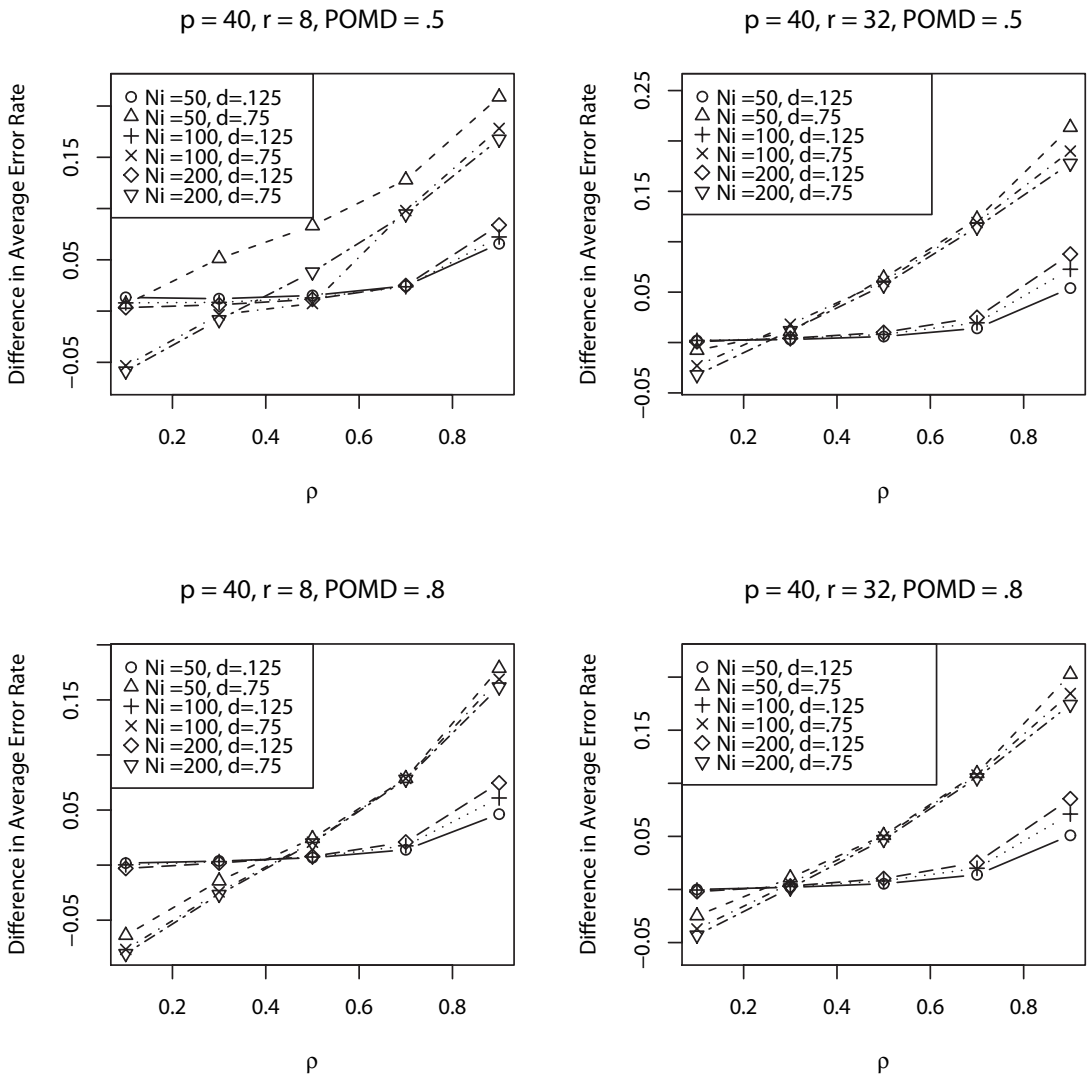

Figure 4.6: Graphs of the $\widehat{EERD}$ versus ρ for fixed values of N_i, r, d_j, $POMD$, and $p = 40$.

where $CER^*\left(W_c^*\right)$ is defined in (4.43). Also, conditioning on $\hat{\boldsymbol{\mu}}_i^*$, $i = 1, 2$, and $\widehat{\boldsymbol{\Sigma}}^*$, the bootstrap CERs for the MLE classifier are

$$CER_{12}^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \widehat{\boldsymbol{\Sigma}}^*\right) \equiv P\left[W_{MLE}^* > 0 | \hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \hat{\boldsymbol{\Sigma}}^*, \mathbf{x} \in \Pi_1\right],$$

where $\mathbf{x}$ is a complete unlabeled observation and W_{MLE}^* is similar in definition to W_{MLE} in (4.36). Given $\mathbf{x} \in \Pi_1$ and $\hat{\delta} \equiv \hat{\boldsymbol{\mu}}_1^* - \hat{\boldsymbol{\mu}}_2^*$, we have

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

and given $\mathbf{x} \in \Pi_2$, we have

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

where

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

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

$$CER^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \widehat{\boldsymbol{\Sigma}}^*\right) = \frac{1}{2}\left[CER_{12}^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \widehat{\boldsymbol{\Sigma}}^*\right) + CER_{21}^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \widehat{\boldsymbol{\Sigma}}^*\right)\right]. \quad (4.44)$$

Then, the estimated bootstrap expected error rate for the MLE classifier is

$$\widehat{EER}_{Boot(MLE)} = \frac{1}{k}\sum_{j=1}^{k} CER_j^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \widehat{\boldsymbol{\Sigma}}^*\right),$$

where

$$CER^*\left(\hat{\boldsymbol{\mu}}_1^*, \hat{\boldsymbol{\mu}}_2^*, \widehat{\boldsymbol{\Sigma}}^*\right)$$

is given in (4.44), k is the total number of simulated training-data sets, and j denotes the j^{th} simulated training-data set where $j \in \{1, 2, ..., k\}$. Therefore, the estimated parametric bootstrap expected error rate difference for the *C-H* and *MLE* classifiers is

$$\widehat{EERD}_{Boot} = \widehat{EER}_{Boot(C\text{-}H)} - \widehat{EER}_{Boot(MLE)}, \quad (4.45)$$

which we use to compare the *C-H* and *MLE* classifiers for two real data sets in the following subsections.

4.4.2 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 Chung and Han (2000). 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 United States student, variables X_1, X_2, X_3, and X_4 are given; however, X_5 contains monotone missing data.

Table 4.1. 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	

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

$$\widehat{\mathbf{C}}_{UTA} = \begin{bmatrix} 1 & 0.145 & -0.066 & 0.199 & 0.373 \\ 0.145 & 1 & -0.404 & 0.494 & 0.767 \\ -0.066 & -0.404 & 1 & 0.129 & -0.493 \\ 0.199 & 0.494 & 0.129 & 1 & 0.392 \\ 0.373 & 0.767 & -0.493 & 0.392 & 1 \end{bmatrix}. \tag{4.46}$$

We remark that only one sample correlation coefficient in (4.46) has a magnitude exceeding 0.50, which reflects relatively low correlation among the features.

First, we verify that we can reasonably assume the data to come from two multivariate normal distributions of the form of (4.7). Tan, Fang, Tian, and Wei (2005) have developed a method for testing the assumption of multivariate normality with small sample sizes using multiple imputations. We use a similar concept based on the Tan et al. (2005) multiple imputation method using Mardia's tests for multivariate skewness and kurtosis.

Specifically, we compute 30 imputations for the missing data and substitute them into the original dataset, resulting in 30 datasets. Next, for each data set we determine p-values for Mardia's tests for multivariate skewness and kurtosis. We then determine the quantiles for the p-values that are summarized in Table 4.2. From Table 4.2, we see that the median p-values for the multivariate skewness and kurtosis test statistics for the data from both Π_1 and Π_2 give no evidence contradicting the assumption that the data for each group follows a multivariate normal distribution. Thus, we assume multivariate normality for both training-data sets given in Table 4.1.

To estimate the $EERD$ for the *C-H* classifier (4.15) and the MLE classifier (4.36) for the UTA Admissions data, we determine the $\widehat{EERD}_{Boot}$ given in (4.45) for the classifiers (4.15) and (4.36) using 10,000 bootstrap simulation iterations, with $p = 5$, $r = 1$, $N_i = 20$, and $n_i = 10$ for $i = 1, 2$. Additionally, the bootstrap mul-

Table 4.2. Quantiles of p-values for Mardia's Tests: UTA Admissions

	Π_1: Success		Π_2: Failure	
Quantile	Skewness	Kurtosis	Skewness	Kurtosis
100 % Max	.903	.393	.794	.990
99 %	.903	.393	.794	.990
95 %	.898	.390	.790	.924
90 %	.850	.311	.779	.964
75 % Q3	.820	.200	.621	.855
50 % Median	.713	.148	.331	.645
25 % Q1	.545	.103	.102	.500
10 %	.497	.090	.036	.412
5 %	.412	.081	.021	.392
1 %	.366	.055	.007	.299
0 % Min	.366	.055	.007	.299

tivariate normal distribution parameters, which are the MLEs for the multivariate normal population parameters given in Theorem 4.1, are

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

and

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

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

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.150 & 6.020 & -2.760 & 8.540 & 6.510 \\ 6.020 & 11504.500 & -4683 & 5859.375 & 3711.097 \\ -2.760 & -4683 & 11701.500 & 1518.625 & -2406.740 \\ 8.540 & 5859.375 & 1518.625 & 12229.187 & 1953.163 \\ 6.510 & 3711.097 & -2406.74 & 1953.163 & 2034.414 \end{bmatrix}.$$

Subsequent to deriving the $\widehat{EERD}_{Boot}$, we obtain $\widehat{EERD}_{Boot} = -0.027$ with the estimated standard error $\left(\widehat{EERD}\right)_{Boot} = 0.001$, which indicates that the C-H procedure yields slightly better discriminatory performance compared to the MLE classifier for the UTA data. The fact that the C-H procedure slightly outperforms the MLE for the UTA data set in terms of $\widehat{EERD}_{Boot}$ is not surprising. In the UTA dataset,

relatively little correlation exists among the features, and the *C-H* method does not require or use information in the correlation between the features with no missing data and those with missing data. However, the MLE classifier does require at least a moderate degree of correlation to estimate the population parameters effectively.

4.4.3 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 *MLE* classifiers is a subset of the well-known Iris data, which is one of the most popular datasets applied in pattern recognition literature and was first analyzed by R. A. Fisher (1936). The data used here is given in Table 4.3. The University of Irvine Machine Learning

Table 4.3. 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	

Repository website provides the original dataset, 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 dataset 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 common estimated correlation matrix is

$$\widehat{\mathbf{C}}_{Iris} = \begin{bmatrix} 1 & 0.716 & 0.708 & 0.549 \\ 0.716 & 1 & 0.473 & 0.651 \\ 0.708 & 0.473 & 1 & 0.677 \\ 0.549 & 0.651 & 0.677 & 1 \end{bmatrix}. \tag{4.47}$$

In (4.47), only one estimated correlation coefficient with magnitude is less than 0.50, which reflects a moderate degree of correlation among the features.

Again, we have utilized Mardia's tests for multivariate skewness and kurtosis to test that we can reasonably assume the data comes from multivariate normal distributions resembling (4.7). We have generated 30 imputation sets for the partial Iris dataset using regression imputation and p-values for Mardia's multivariate skewness and kurtosis test statistics which are given with the corresponding quantiles in Table 4.4. Clearly, the median p-values for Mardia's multivariate skewness and kurtosis statistics for both Π_1 and Π_2 give no statistical evidence contradicting the assumption that the data for each group follows a multivariate normal distribution.

As in the bootstrap estimated EER comparison for the UTA data, we have used 10,000 bootstrap simulation iterations for calculating the $\widehat{EERD}_{Boot}$ for the Iris subset data. Here, for the Iris data given in Table 4.3, $r = 1$, $N_i = 20$, and $n_i = 10$ for $i = 1, 2$, and $p = 4$. The bootstrap parameters corresponding to Π_1 and Π_2 are

$$\hat{\boldsymbol{\mu}}_1 = [5.035, 3.48, 1.435, .235]'$$

Table 4.4. Quantiles of p-values for Mardia's Tests: Partial Iris Data

	Π_1: Setosa		Π_2: Versicolor	
Quantile	Skewness	Kurtosis	Skewness	Kurtosis
100 % Max	.987	.981	.914	.828
99 %	.987	.981	.914	.828
95 %	.912	.956	.884	.638
90 %	.886	.921	.840	.546
75 % Q3	.822	.848	.785	.401
50 % Median	.575	.752	.638	.247
25 % Q1	.334	.573	.522	.192
10 %	.088	.433	.285	.134
5 %	.060	.369	.035	.123
1 %	.032	.290	.001	.108
0 % Min	.032	.290	.001	.108

and

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

respectively, with common covariance matrix

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

Comparing our parametric bootstrap estimates for the *C-H* and *MLE* classifiers applied to the subset of the Iris dataset, we obtain $\widehat{EERD}_{Boot} = 0.11$ with an estimated standard error $\widehat{EERD}_{Boot} = 0.001$, which indicates that $\widehat{EER}_{Boot(MLE)} \ll \widehat{EER}_{Boot(C\text{-}H)}$. In (4.47), we see that the sample correlation coefficients r_{14}, r_{24}, and r_{34} are all at least moderately large, which demonstrates moderately high correlation for X_1, X_2, and X_3 in the partial Iris data with X_4, the feature that contains observations with monotone missing elements. Consequently, the MLE classifier convincingly outperforms the *C-H* classifier in terms of the bootstrap $\widehat{EERD}_{Boot}$.

Thus, comparing *C-H* and *MLE* classifiers on real data, we have provided additional evidence that, indeed, the MLE classifier incorporates more information

into the discriminatory function by effectively utilizing the correlation among the variables with missing and non-missing observations provided a sufficient degree of correlation exists. Hence, from our two Monte Carlo simulation studies and our estimates of $\widehat{EERD}_{Boot}$ for two real data sets, we see that for the Monte Carlo simulation parameter configurations, sample sizes, and the real-data sets considered here, the MLE classifier is preferred to the *C-H* classifier for monotone missing data with equal covariance matrices where the correlation between features with no missing data with features having monotone missing data is moderate to large. This evidence essentially contradicts the simulation results in Chung and Han (2000) that the *C-H* classifier is superior to the MLE classifier of Hocking and Smith (1968).

4.5 Conclusion

In this chapter, we have considered the problem of supervised classification using training data with identical monotone missing data patterns for two classes. Subsequently, we have introduced the *C-H* classifier for monotone missing data from Chung and Han (2000), which has yielded a classifier composed of a linear combination of the complete and incomplete data. Moreover, we have specified the overall CERs and EERs for the *C-H* classifier and have derived the MLEs for the partitioned population class means and the common covariance matrix. Furthermore, we have derived an expression for the MLE linear classifier when monotone missing training data is present in both training-data sets. Also, we have stated the corresponding overall CERs and EERs for the MLE classifier. We have then used two Monte Carlo simulations to demonstrate that for the various parameter configurations considered here, ρ and d have the greatest impact on $\widehat{EERD}$. We also have concluded that if $\rho \geq .5$, the MLE classifier becomes an increasingly superior statistical classification procedure as ρ approaches 1. This conclusion essentially contradicts the simulation results of Chung and Han (2000).

We also have compared the MLE and C-H classifiers on two real training-data sets using $\widehat{EERD}_{Boot}$ defined in (4.45). From the data set from Chung and Han (2000), we have demonstrated that when the correlation among features without missing data and the feature with missing data is small, the C-H classifier is slightly better than the MLE classifier ($\widehat{EERD}_{Boot} = -0027$). Finally, we have used a subset of the prominent Iris data set from Fisher (1936) to illustrate that when correlation among features without missing data and features with missing data is moderate to large, the MLE classifier is conclusively better than the C-H classifier ($\widehat{EERD}_{Boot} = 0.11$).

APPENDIX

APPENDIX A

Chapter Four

Lemma. Let

$$\mathbf{W}_i = \mathbf{B}_{i1} + \mathbf{B}_{i2}, \tag{A.1}$$

with

$$\mathbf{B}_{i1} = \sum_{j=n_i+1}^{N_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2})'$$

and

$$\mathbf{B}_{i2} = \frac{n_i (N_i - n_i)}{N_i} (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_{i2}) (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_{i2})',$$

where $\bar{\mathbf{y}}_i$, $\bar{\mathbf{y}}_{i1}$, and $\bar{\mathbf{y}}_{i2}$ are defined in (4.11), (4.12), and (4.13), respectively. Also, let $\mathbf{A}_{11,n_i,i}$ and $\mathbf{A}_{11,N_i,i}$ be defined as in (4.28) and (4.27), respectively. Then

$$\mathbf{W}_i = \mathbf{A}_{11,N_i,i} - \mathbf{A}_{11,n_i,i}. \tag{A.2}$$

Proof: First, note that $\bar{\mathbf{y}}_i = \frac{n_i}{N_i}\bar{\mathbf{y}}_{i1} + \frac{(N_i-n_i)}{N_i}\bar{\mathbf{y}}_{i2}$, $\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i = \frac{(N_i-n_i)}{N_i} (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_{i2})$, and $\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_i = -\frac{n_i}{N_i} (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_{i2})$. Hence,

$$\begin{aligned}
\mathbf{B}_{i2} &= \frac{n_i (N_i - n_i)}{N_i} (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i + \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{i2}) (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i + \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_{i2})' \\
&= n_i (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i) (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i)' + (N_i - n_i) (\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_i) (\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_i)' \\
&= \sum_{j=1}^{n_i} (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i) (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_i)' + \sum_{j=n_i+1}^{N_i} (\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_i) (\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_i)'.
\end{aligned}$$

Next,

$$
\begin{aligned}
&\mathbf{A}_{11,n_i} + \mathbf{B}_{i1} + \mathbf{B}_{i2} \\
&\quad = \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i1}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i1})' + \sum_{j=n_i+1}^{N_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2})' \\
&\qquad + \sum_{j=1}^{n_i} (\mathbf{y}_{i1} - \bar{\mathbf{y}}_{i}) (\mathbf{y}_{i1} - \bar{\mathbf{y}}_{i})' + \sum_{j=n_i+1}^{N_i} (\mathbf{y}_{i2} - \bar{\mathbf{y}}_{i}) (\mathbf{y}_{i2} - \bar{\mathbf{y}}_{i})' \\
&\quad = \sum_{j=1}^{n_i} \big[(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i1}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i1})' + (\mathbf{y}_{i1} - \bar{\mathbf{y}}_{i}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i1})' + (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_{i}) (\bar{\mathbf{y}}_{i1} - \bar{\mathbf{y}}_{i})' \\
&\qquad + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i1}) (\mathbf{y}_{i1} - \bar{\mathbf{y}}_{i})' \big] \\
&\qquad + \sum_{j=n_i+1}^{N_i} \big[(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2})' + (\mathbf{y}_{i2} - \bar{\mathbf{y}}_{i}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2})' \\
&\qquad + (\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_{i}) (\bar{\mathbf{y}}_{i2} - \bar{\mathbf{y}}_{i})' + (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i2}) (\mathbf{y}_{i2} - \bar{\mathbf{y}}_{i})' \big] \\
&\quad = \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i})' + \sum_{j=n_i+1}^{N_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i}) (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i})' \\
&\quad = \mathbf{A}_{11,N_i,i}.
\end{aligned}
$$

Therefore, (A.2) holds.

Proof of Theorem 4.1

Proof: Let

$$
\mathbf{A}_i \equiv \begin{bmatrix} \mathbf{A}_{11,n_i,i} & \mathbf{A}_{12,n_i,i} \\ \mathbf{A}_{21,n_i,i} & \mathbf{A}_{22,n_i,i} \end{bmatrix}, \tag{A.3}
$$

where $\mathbf{A}_{11,n_i,i}$, $\mathbf{A}_{12,n_i,i}$, and $\mathbf{A}_{22,n_i,i}$ are defined in (4.28), (4.29), and (4.30), respectively, and recall that $\mathbf{A}_i \in \mathbb{R}_{p\times p}$. Also, let $\mathbf{W}_i \in \mathbb{R}_{k\times k}$ be defined in (A.1), and let $\mathbf{\Sigma} \in \mathbb{R}^{>}_{p\times p}$ be defined as in (4.26). Following Anderson and Olkin (1985), the

concentrated log likelihood for $\mathbf{\Sigma}$ is

$$\ln L\left(\mathbf{\Sigma}|all\ data\right) = \sum_{i=1}^{2}\left[-\left(\frac{N_i-n_i}{2}\right)|\mathbf{\Sigma}_{11}| - \left(\frac{n_i}{2}\right)|\mathbf{\Sigma}|\,\frac{1}{2}tr\left(\mathbf{\Sigma}^{-1}\mathbf{A}_i\right) + \frac{1}{2}tr\left(\mathbf{\Sigma}_{11}^{-1}\mathbf{W}_i\mathbf{\Sigma}_{11}^{-1}\right)\right].$$

Therefore,

$$\frac{\partial \ln L}{\partial \mathbf{\Sigma}} = \sum_{i=1}^{2}\left(-\left(\frac{n_i}{2}\right)\mathbf{\Sigma}^{-1} + \frac{1}{2}\left(\mathbf{\Sigma}^{-1}\mathbf{A}_i\mathbf{\Sigma}^{-1}\right) - \left(\frac{N_i-n_i}{2}\right)\begin{bmatrix}\mathbf{\Sigma}_{11}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} + \frac{1}{2}\begin{bmatrix}\mathbf{\Sigma}_{11}^{-1}\mathbf{W}_i\mathbf{\Sigma}_{11}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right). \tag{A.4}$$

Setting (A.4) to zero and multiplying both sides of (A.4) by $\mathbf{\Sigma}$ and 2 implies

$$\sum_{i=1}^{2}\left(-n_i\begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22}\end{bmatrix} - (N_i-n_i)\begin{bmatrix}\mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\end{bmatrix} + \begin{bmatrix}\mathbf{A}_{11,n_i,i} & \mathbf{A}_{12,n_i,i}\\ \mathbf{A}_{21,n_i,i} & \mathbf{A}_{22,n_i,i}\end{bmatrix} + \begin{bmatrix}\mathbf{W}_i & \mathbf{W}_i\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\\ \mathbf{W}_i\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{W}_i\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\end{bmatrix}\right) = \mathbf{0}.$$

Hence, we obtain the three estimating equations

$$\sum_{i=1}^{2}\left[-n_i\mathbf{\Sigma}_{11} + \mathbf{A}_{11,n_i,i} - (N_i-n_i)\,\mathbf{\Sigma}_{11} + \mathbf{W}_i\right] = \mathbf{0}, \tag{A.5}$$

$$\sum_{i=1}^{2}\left[-(N_i-n_i)\,\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{W}_i\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} - n_i\mathbf{\Sigma}_{22} + \mathbf{A}_{22,n_i,i}\right] = \mathbf{0}, \tag{A.6}$$

and

$$\sum_{i=1}^{2}\left[-(N_i-n_i)\,\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{W}_i\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} - n_i\mathbf{\Sigma}_{22} + \mathbf{A}_{22,n_i,i}\right] = \mathbf{0}. \tag{A.7}$$

Solving (A.5) for $\mathbf{\Sigma}_{11}$, we get

$$\begin{aligned}
&\sum_{i=1}^{2}[-n_i\mathbf{\Sigma}_{11} + \mathbf{A}_{11,n_i,i} - (N_i - n_i)\,\mathbf{\Sigma}_{11} + \mathbf{W}_i] = \mathbf{0} \\
&\quad\Rightarrow \left(\sum_{i=1}^{2} N_i\right)\mathbf{\Sigma}_{11} = \sum_{i=1}^{2}\mathbf{A}_{11,n_i,i} + \sum_{i=1}^{2}[\mathbf{A}_{11,N_i,i} - \mathbf{A}_{11,n_i,i}] \\
&\quad\Rightarrow \widehat{\mathbf{\Sigma}}_{11} = \frac{\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i}}{\sum_{i=1}^{2} N_i}.
\end{aligned}$$

Next, solving (A.6) for $\mathbf{\Sigma}_{12}$, we get

$$\begin{aligned}
&\sum_{i=1}^{2}\left[-N_i\mathbf{\Sigma}_{12} + \mathbf{A}_{12,n_i,i} - (N_i - n_i)\,\mathbf{\Sigma}_{12} + [\mathbf{A}_{11,N_i,i} - \mathbf{A}_{11,n_i,i}]\,\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\right] = \mathbf{0} \\
&\quad\Rightarrow \sum_{i=1}^{2}\mathbf{A}_{12,n_i,i} = \left(\sum_{i=1}^{2} N_i\right)\mathbf{\Sigma}_{12} - \left(\sum_{i=1}^{2} N_i\right)\left[\sum_{i=1}^{2}[\mathbf{A}_{11,N_i,i} - \mathbf{A}_{11,n_i,i}]\right]\times \\
&\qquad\qquad \left[\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i}\right]^{-1}\mathbf{\Sigma}_{12} \\
&\quad\Rightarrow \sum_{i=1}^{2}\mathbf{A}_{12,n_i,i} = \left(\sum_{i=1}^{2} N_i\right)\left[\mathbf{I} - \left[\sum_{i=1}^{2}[\mathbf{A}_{11,N_i,i} - \mathbf{A}_{11,n_i,i}]\right]\left[\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i}\right]^{-1}\right]\mathbf{\Sigma}_{12} \\
&\quad\Rightarrow \sum_{i=1}^{2}\mathbf{A}_{12,n_i,i} = \left(\sum_{i=1}^{2} N_i\right)\left(\left[\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i} - \sum_{i=1}^{2}[\mathbf{A}_{11,N_i,i} - \mathbf{A}_{11,n_i,i}]\right]\right. \\
&\qquad\qquad \left.\left[\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i}\right]^{-1}\right)\mathbf{\Sigma}_{12} \\
&\quad\Rightarrow \sum_{i=1}^{2}\mathbf{A}_{12,n_i,i} = \left(\sum_{i=1}^{2} N_i\right)\left[\sum_{i=1}^{2}\mathbf{A}_{11,n_i,i}\right]\left[\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i}\right]^{-1}\mathbf{\Sigma}_{12} \\
&\quad\Rightarrow \widehat{\mathbf{\Sigma}}_{12} = \frac{1}{\left(\sum_{i=1}^{2} N_i\right)}\left[\sum_{i=1}^{2}\mathbf{A}_{11,N_i,i}\right]\left[\sum_{i=1}^{2}\mathbf{A}_{11,n_i,i}\right]^{-1}\left[\sum_{i=1}^{2}\mathbf{A}_{12,n_i,i}\right].
\end{aligned}$$

Finally, solving (A) for $\mathbf{\Sigma}_{22}$, we get

$$\sum_{i=1}^{2}\left[-(N_i - n_i)\,\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{A}_{11,N-n,i}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} - n_i\mathbf{\Sigma}_{22} + \mathbf{A}_{22,n,i}\right] = \mathbf{0},$$

which implies

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

$$
\begin{aligned}
&= \sum_{i=1}^{2} \mathbf{A}_{22,n_i,i} + \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \times \\
&\quad \left[\frac{\sum_{i=1}^{2} N_i}{\sum_{i=1}^{2} N_i} \sum_{i=1}^{2} \mathbf{A}_{11,N_i,i} - \sum_{i=1}^{2} \mathbf{A}_{11,n_i,i} - \frac{\sum_{i=1}^{2} (N_i - n_i)}{\sum_{i=1}^{2} N_i} \sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}\right] \times \\
&\quad \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{2} \mathbf{A}_{22,n_i,i} - \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right] \times \\
&\quad \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right] \left(\frac{\sum_{i=1}^{2} n_i}{\sum_{i=1}^{2} N_i}\right) \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \times \\
&\quad \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{2} \mathbf{A}_{22,n_i,i} - \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right] \\
&\quad + \left(\frac{\sum_{i=1}^{2} n_i}{\sum_{i=1}^{2} N_i}\right) \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}\right] \times \\
&\quad \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right]
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \widehat{\mathbf{\Sigma}}_{22} =& \frac{1}{\sum_{i=1}^{2} n_i} \sum_{i=1}^{2} \mathbf{A}_{22\cdot 1,n_i,i} + \frac{1}{\sum_{i=1}^{2} N_i} \left[\sum_{i=1}^{2} \mathbf{A}_{21,n_i,i}\right] \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{11,N_i,i}\right] \times \\
& \left[\sum_{i=1}^{2} \mathbf{A}_{11,n_i,i}\right]^{-1} \left[\sum_{i=1}^{2} \mathbf{A}_{12,n_i,i}\right],
\end{aligned}
$$

where

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

The derivation of the *MLE*s for $\boldsymbol{\mu}_{i1}$ and $\boldsymbol{\mu}_{i2}$, $i = 1, 2$, for two multivariate normal populations using identical monotone missing data patterns for both samples is similar to the derivation of the *MLE* of

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$$

for a single multivariate normal population with monotone missing data given in Anderson (1957).

BIBLIOGRAPHY

Anderson, T. W. (1951), "Classification by multivariate analysis," *Psychometrika*, 16, 31 – 50.

— (1957), "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing," *Journal of the American Statistical Association*, 278, 200 – 203.

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.

Arnold, B. C. and Beaver, R. J. (2002), "Skewed multivariate models related to hidden truncation and/or selective reporting," *Sociedad de Estadistica e Investigacion Operativa*, 11, 7 – 54.

Azzalini, A. (1985), "A class of distributions which include the normal ones," *Scand. J. Statist.*, 12, 171 – 178.

Azzalini, A. and Capitanio, A. (1999), "Statistical applications of the multivariate skew normal distribution," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61, 579 – 602.

Balakrishnama, S. and Ganapathiraju, A. (1998), "Linear discriminant analysis - a brief tutorial," Tech. rep., Institute for Signal and Information Processing, Department of Electrical and Computer Engineering, Mississippi State University.

Batsidis, A., Zografos, K., and Loukas, S. (2006), "Errors in discrimination with monotone missing data from multivariate normal populations," *Computational Statistics & Data Analysis*, 50, 2600 – 2634.

Bellman, R. (1961), *Adaptive Control Processes: A Guided Tour*, Princeton University Press.

Bohannon, T. R. and Smith, W. B. (1975), "Classification based on incomplete data records," *ASA Proc. Social Statistics Section*, 67, 214 – 218.

Carreira-Perpinan, M. A. (1997), "A review of dimension reduction techniques," Tech. rep., University of Sheffield.

Chan, L., Gilman, A., and Dunn, O. J. (1976), "Alternative approaches to missing values in discriminant analysis," *Journal of the American Statistical Association*, 71, 842 – 844.

Chan, 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.

Chen, J. (2007), "Theoretical results and applications related to dimension reduction," Master's thesis, Georgia Institute of Technology.

Chung, H.-C. and Han, C.-P. (2000), "Discriminant analysis when a block of observations is missing," *Ann. Inst. Statist. Math.*, 52, 544 – 556.

Cunningham, P. (2007), "Dimension reduction," Tech. rep., University College Dublin.

Decell, H. P., Odell, P. L., and Coberly, W. A. (1981), "Linear dimension reduction and Bayes classification," *Pattern Recognition*, 13, 242 – 243.

Donoho, D. L. (2000), "High-dimensional data: the curses and blessings of dimensionality," .

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals Eugenics*, 7, 179 – 188.

Fodor, I. K. (2002), "A survey of dimension reduction techniques," Tech. rep., Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

Friedman, J. H. and Tukey, J. W. (1974), "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers C-23*, 9, 881 – 890.

Goodman, I. R. and Kotz, S. (1973), "Multivariate θ-generalized normal distributions," *Journal of Multivariate Analysis*, 3, 204 – 219.

Hamsici, O. C. and Martinez, A. (2008), "Bayes optimality in linear discriminant analysis," *IEEE Transactions on Pattern Analysis and Machine Learning*, 30, 647 – 657.

Hennig, C. (2004), "Asymmetric linear dimension reduction for classification," *J. Comput. Graph. Statist.*, 13, 930–945.

Hocking, R. R. and Smith, W. B. (1968), "Estimation of parameters in the multivariate normal distribution with missing observations," *Journal of the American Statistical Association*, 63, 159 – 173.

Jackson, E. C. (1968), "Missing values in linear multiple discriminant analysis," *Biometrics*, 24, 835 – 844.

John, S. (1961), "Errors in discrimination," *Ann. Math. Statist.*, 32, 1125 – 1144.

Khambatla, N. and Leen, T. K. (1997), "Dimension reduction by local principal component analysis," *Neural Computation*, 7, 1493–1516.

Khatri, C. G. (1968), "Some results for the singular normal multivariate regression models," *Sankhya A*, 30, 267 – 280.

Lewis, T. O. and Odell, P. L. (1971), *Estimation in Linear Models*, Prentice-Hall.

Loog, M. and Duin, R. P. W. (2001), "Multiclass linear dimension reduction by weighted pairwise fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, 23, 762–766.

— (2004), "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Learning*, 26, 732–739.

Lotikar, R. and Kothari, R. (2000), "Adaptive linear dimensionality reduction for classification," *Pattern Recognition*, 33, 185–194.

Martinez, A. M. and Kak, A. C. (2001), "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 228 – 233.

Odell, P. L. (1979), "A model for dimension reduction in pattern recognition using continuous data," *Pattern Recognition*, 11, 51 – 54.

Onsupreth, S. T. and Young, D. M. (2005), "A note on discriminant analysis with the multivariate normal distribution," Tech. rep., Baylor University.

Pavlenko, T. (2003), "On feature selection, curse-of-dimensionality and error probability in discriminant analysis," *Journal of Statistical Planning and Inference*, 115, 565 – 584.

Peters, C., Redner, R., and Decell, H. P. (1978), "Characterizations of linear sufficient statistics," *Sankhya A*, 40, 303 – 309.

Rao, C. R. (1948), "The utilization of multiple measurements in problems of biological classification," *J. Roy. Statist. Soc. B*, 10, 159 – 203.

Ravisekar, B. (2006), "A comparative analysis of dimensionality reduction techniques," Tech. rep., College of Computing Georgia Institute of Technology.

Tan, M., Fang, H.-B., Tian, G.-L., and Wei, G. (2005), "Testing multivariate normality in incomplete data of small sample size," *Journal of Multivariate Analysis*, 93, 164 – 179.

Tang, E. K., Suganthan, P. N., Yao, X., and Qin, A. K. (2005), "Linear dimensionality reduction using relevance weighted LDA," *Pattern Recognition*, 38, 485–493.

Titterington, D. M. and Jiang, J.-M. (1983), "Recursive estimation procedures for missing-data problems," *Biometrika Trust*, 70, 613 – 624.

Tubbs, J. D., Coberly, W. A., and Young, D. M. (1982), "Linear dimension reduction and Bayes classification with unknown population parameters," *Pattern Recognition*, 4, 243 – 255.

van Perlo-ten Kleij, F. (2004), "Contributions to Multivariate Analysis with Applications in Marketing," Ph.D. thesis, University of Groningen.

Vernic, R. (2005), "On the multivariate skew-normal distribution and its scale mixtures," *A. Stiint. Univ. "Ovidius" Constanta, Ser. Math.*, 13, 83 – 96.

— (2006), "Multivariate skew-normal distributions with applications in insurance," *Insurance: Mathematics and Economics*, 38, 413 – 426.

Warner, R. M. (2007), *Applied Statistics: From Bivariate Through Multivariate Techniques*, Sage Publications, Inc.

Young, D. M., Odell, P. L., and Marco, V. R. (1985), "Optimal linear feature selection for a general class of statistical pattern recognition models," *Pattern Recognition Recognition Letters*, 3, 161 – 165.