COVARIATE ADJUSTED DISCRIMINATION WITH APPLICATIONS TO NEUROSCIENCE

by

Josephine Asafu-Adjei

B.S. in Math/Economics, University of Pittsburgh, 2004

M.A. in Applied Statistics, University of Pittsburgh, 2007

Submitted to the Graduate Faculty of

the Kenneth P. Dietrich School of Arts & Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2011

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS & SCIENCES

This dissertation was presented

by

Josephine Asafu-Adjei

It was defended on

August 29, 2011

and approved by

Dr. Allan R. Sampson

Dr. Leon J. Gleser

Dr. Satish Iyengar

Dr. Chien-Cheng Tseng

Dissertation Director: Dr. Allan R. Sampson

COVARIATE ADJUSTED DISCRIMINATION WITH APPLICATIONS TO NEUROSCIENCE

Josephine Asafu-Adjei, PhD

University of Pittsburgh, 2011

In post-mortem tissue studies that compare regional brain biomarkers across different mental disorder diagnostic groups, subjects are often matched on several demographic characteristics and measured on additional covariates. The goal of our research is to integrate the results from these types of studies using two commonly used statistical discrimination techniques, namely, linear discriminant analysis (LDA) and classification trees based on the algorithm developed by Breiman, Friedman, Olshen, and Stone (BFOS), to identify the most discriminatory subset of biomarkers. Subject matching and covariate effects don't appear in the literature implementing these discriminatory methods in the analysis of post-mortem tissue studies (e.g., Knable et al. 2001; Knable et al. 2002).

Although there are methods that have been developed for LDA to account for covariate effects on the response or feature variables of interest, none of these methods addresses the fact that individuals may also be matched across several groups. One aspect of our research extends this work to handle group matching.

To develop the theoretical foundations required to account for covariate effects in classification trees, we describe how to implement the BFOS algorithm, which is non-parametric and traditionally implemented in a data based setting, when the feature variables come from a known distribution. We then extend this algorithm to the case where the feature variables come from a known distribution, conditional on a covariate value. From this development, we carefully formulate a semi-parametric model for the conditional distribution of the feature

variables that allows the use of the BFOS algorithm to construct a covariate adjusted tree based on one unique set of feature variables, in both a theoretical setting and in the context of training data. Finally, the tree construction procedure we develop using this conditional model is extended to handle group matching.

Our adjustment methodology is successfully applied to a series of post-mortem tissue studies conducted by Sweet et al. (2003, 2004, 2007, 2008) comparing several neurobiological characteristics of schizophrenia subjects and normal controls, and to a post-mortem tissue study conducted by Konopaske et al. (2008) comparing brain biomarker measurements of monkeys across three treatment groups.

Keywords: linear discriminant analysis, classification trees, recursive partitioning algorithm, matched design, post-mortem tissue studies, schizophrenia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would first like to express my sincere gratitude to the entire faculty and staff in the Statistics department for all of their help, support, and guidance throughout my six years in the graduate program. However, I would like to personally acknowledge several people who have been especially instrumental in my personal growth and development during my time in the program. First and foremost, I would like to thank my advisor, Dr. Sampson, for his mentoring, guidance, research support, and endless kindness and patience. Most of all, I would like to thank him for his unwavering confidence in my abilities and for always encouraging me to do better. Without his confidence and faith in me, I would not have attained the level of success in my graduate student career that I have achieved.

I am also indebted to my committee members for taking the time to look through my dissertation and give me feedback on areas in which to improve. With their assistance, the quality of my dissertation has greatly improved. I am grateful to Dr. Tseng for his constructive comments and suggestions regarding my dissertation. I would also like to thank Drs. Gleser and Iyengar not just for being my committee members, but also for being excellent educators to me over the years. I've looked to them as my mentors and I can't say enough how much I appreciate all of their advice, guidance, and support.

I am especially grateful to Dr. Bodenschatz for seeing the potential I had as an undergraduate to flourish as a statistician and for encouraging me to enroll in the Statistics graduate program in the first place. I also appreciate all the help he's given me ever since. I would like to thank Drs. Pfenning, Krafty, Cheng, and Block for all the kindness and generosity they've shown me over the years. In addition, I'm very grateful to Dr. Krafty for taking the time to give me invaluable career advice. I would also like to thank Mary Gerber and Kim Thomas for their immeasurable help, kindness, and support.

Finally, I would like to give a special thank you to my parents for all of their love and

support, and for all the sacrifices they have made over the years to get me to where I am today. Thank you to my sisters Nana, Abena, and Barbara, and my brother Kofi for all of their loyalty, encouragement, and for being the best siblings one could ever have. Thank you to Chen, Rose, Chioma, and Scott for being such kind, generous, and loyal friends. Last, but certainly not least, I want to give another special thank you to my husband Nana for his unwavering kindness, support, and encouragement during my time in the program. I love you all, God bless.

1.0 INTRODUCTION

1.1 BACKGROUND

In the statistical analysis of a particular set of response or feature variables measured on an individual, it is possible that an individual on whom these variables are measured may belong to one of g ($g \geq 2$) groups. In this case, if is often of interest to determine which of these feature variables best differentiates among individuals belonging to these groups. The motivation of our research and its ultimate application is on the analysis of post-mortem brain tissue studies, which are used in neuroscience to detect differences in regional brain biomarker measurements between subjects from different mental disorder diagnostic groups, e.g., normal controls and subjects with schizophrenia. Over time, an increasing number of these studies have been done on the same cohort of subjects, where each study considers different biomarkers. It is of considerable interest to integrate the evidence from a set of such comparative post-mortem tissue studies in order to identify among the examined biomarkers those that best discriminate among the diagnostic groups under consideration. In general, the identification of subsets of discriminatory biomarkers in such studies tends to be more exploratory in nature with a goal to obtain better characterizations of the pathology or pathologies of interest. The insights gained can be used in developing new hypotheses that can be tested prospectively.

Rather than just fitting univariate models to each feature variable to determine which of them significantly differs with respect to group, discrimination approaches obtain the most discriminatory subset of feature variables by taking into account the interrelationships that exist among these variables. Among the various statistical methods that accomplish this, two are commonly used in practice, namely, linear discriminant analysis (LDA) and classifi-

cation trees. Linear discriminant analysis, first introduced by R. A. Fisher[8][9], is based on the assumption that the feature data follow a normal distribution with a common variance-covariance matrix across groups, where this latter assumption has been relaxed since Fisher's initial development. Classification trees are constructed using a computationally intensive recursive partitioning algorithm that, unlike LDA, makes no assumptions regarding the distribution of the feature data. One notable complication in using any of these statistical procedures to discriminate among groups occurs when the individuals on whom these feature data are measured are matched across groups on the basis of certain attributes, such as age or gender. In many post-mortem brain tissue studies, individuals from each of the diagnostic groups under consideration are matched to better control the inherent experimental variability that arises due to the manner in which brain tissue is processed. A further challenge in these analyses is to also account for other subject characteristics or covariates not included in the matching process. Although such covariates are not considered germane in differentiating among the groups of interest, they may still have an important impact on the feature variables under consideration.

Illustrative of such integrative analyses aimed at group discrimination are two recent studies conducted by Knable et al. in 2001 [16] and 2002 [15], which are based on post-mortem tissue specimens taken from the Stanley Foundation Neuropathology Consortium. The principal purpose of the Knable et al. studies was to determine a subset of prefrontal cortical markers that best discriminated among the following four diagnostic groups: schizophrenia, bipolar disorder, major depressive disorder (MDD) without psychotic features, and normal controls. In each of these two studies, there were 15 matched quadruples of individuals, one from each of the four groups, where the matching was based on several characteristics, including age at death and post-mortem interval (PMI), which is the amount of elapsed time between actual time of death and time of tissue collection, so that there were a total of 60 subjects in each study. Also, while not matched for brain tissue storage time, the amount of time for which brain tissue has been stored, this covariate was also measured for each subject. Their 2001 study first used a stepwise variant of LDA to determine the most discriminatory subset of prefrontal cortical markers, which subsequently served as a basis for traditional LDA to measure the extent to which this subset correctly classified new individuals belonging to one of these four diagnostic groups. In their 2002 study, the BFOS classification tree

construction algorithm (see Breiman, Friedman, Olshen, and Stone (BFOS)[7]) was used to identify the subset of cortical markers that best distinguished among the four diagnostic groups and measure classification accuracy.

However, Knable et al. did not account for either the subject matching that was used or the measurement of additional covariates, such as brain tissue storage time, in their discriminatory analyses. This omission is potentially problematic due to the fact that cohort processing and covariates can potentially have considerable influence on biomarker measurements. In particular, tissue processing plays an important role in the variability of biomarker measurements across cohorts. Tu et al. also point out that failure to account for design and covariate effects on the feature data may produce misleading results with poor discriminatory ability [37]. In general, the statistical methodology we develop aims to adjust, or control, for the effects of subject matching and covariates in the identification of feature variables that best discriminate among the groups of interest.

Schizophrenia is a chronic, severe, and debilitating mental disorder, characterized mainly by cognition impairment. The Conte Center for the Neuroscience of Mental Disorders (CCNMD) in the Department of Psychiatry at the University of Pittsburgh has been involved in conducting extensive neurobiological research concerning this disorder. One area of research that the Center focuses on is the analyses of post-mortem tissue samples to detect neurobiological abnormalities in subjects with schizophrenia as compared to normal controls. In each of the human post-mortem tissue studies conducted in the Center, every subject with schizophrenia is pair matched with a control subject based on age at death, gender, and PMI, with some studies including an additional matched diagnostic group, e.g., subjects with MDD. Auxiliary covariate data, such as brain pH value and brain tissue storage time, are also collected for each subject. Our goal is to take subject pairing and covariates into account when integrating data from these post-mortem tissue studies in order to accurately determine which biomarkers best discriminate schizophrenia subjects from normal controls.

We reiterate that our interest is primarily focused on discrimination and not classification. Conceptually, we want to answer questions similar to that posed in the following scenario. Suppose one is considering a hypothetical pair in a post-mortem tissue study consisting of a control subject and a subject with schizophrenia who have the same age at death, gender, and PMI and whose measured multiple biomarkers are obtained under the "same

conditions", meaning that both members of the pair had their biomarkers measured in the same manner. This is in recognition that differing batches of the same reagent might vary in strength and, thus, impact the measurement process. The question then becomes which biomarkers best distinguish the subject with schizophrenia from the control subject in any given pair. In doing this discrimination, we also want to take into account the effects of other covariates, such as brain tissue storage time, that were not considered in the pairing. Moreover, we would like these obtained discriminatory biomarkers not to depend on either the characteristics specific to that pair or how that pair was processed.

Although we present our adjustment methodology in the context of post-mortem tissue studies, the applicability of this methodology extends to a wide variety of studies, including traditional epidemiological case-control studies, imaging studies, and genomic studies.

1.2 RESEARCH OVERVIEW

Our research is centered on controlling for the effects of subject matching and additional covariates when determining the discriminatory ability of a particular set of feature variables and classifying new individuals using LDA and classification trees constructed using the BFOS recursive partitioning algorithm.

An overview of post-mortem tissue studies is provided in Chapter 2, followed by a description of the standard statistical models used in these studies. We then discuss one post-mortem tissue data set that, in part, motivates our subsequent research, along with another data set to which we apply our adjustment methodology.

In Chapter 3, we give an overview of classification and consider traditional LDA. A review of covariance adjusted linear discriminant analysis, a modification of traditional LDA that utilizes the conditional distribution of the feature variables of interest, and a description of the relevant literature is subsequently provided. We then introduce the formulation of our method of paired LDA, which extends the methodology developed by Lachenbruch [19] and Tu et al. [37] to handle the case where individuals are paired, as well as the case where individuals are paired and also measured on covariates not included in the pairing. Finally, we extend our adjustment procedure to handle matching across multiple groups.

Chapter 4 begins with an overview of classification trees constructed using the BFOS recursive partitioning algorithm, which is typically used in the context of training data. This is followed by a more detailed description of the classification tree construction procedure, first in a population setting and then in a data setting. We then discuss our modification to the BFOS algorithm that we develop to adjust for the effects of covariates on the feature data. Next, we extend this adjusted recursive partitioning method to develop semi-parametric classification trees, which arise from our assumption that the conditional distribution of the feature data is based on a parametric function of fixed covariate values. We then describe how the procedure used in constructing semi-parametric classification trees can be applied to adjust for the effect of subject matching across two or more groups, along with the effects of additional covariates, on the feature data.

Our adjustment methodology is first applied in Chapter 5 to the analysis of six auditory cortical biomarkers measured in four post-mortem tissue studies conducted by Sweet et al. [33][34][35][36], which compared subjects with schizophrenia with control subjects. We then discuss the application of our methodology to six biomarkers measured in one post-mortem brain tissue study conducted by Konopaske et al. [17] that compared monkeys that were each treated with one of three different drugs, namely, a sham drug, haloperidol, or olanzapine, the latter two of which are antipsychotics.

Finally, we present in Chapter 6 a further discussion of our present research, including the future work we plan to pursue, which includes an extension of our research methodology to quadratic and logistic discriminant analysis, as well as to the tree ensemble construction algorithm of random forests.

2.0 MOTIVATING DATA

2.1 OVERVIEW OF POST-MORTEM TISSUE STUDIES

In the CCNMD, as of May 26, 2011, there are 86 subjects with schizophrenia and 181 control subjects in the Brain Tissue Bank. Post-mortem psychiatric information, such as drug usage and cause of death, has been collected for these subjects, along with demographic information. These subjects have been used repeatedly in studies conducted under the auspices of the CCNMD. In a single study focused on a few select biomarkers, tissue is first obtained for each subject from a specific brain region, such as the primary auditory cortex, and several sections are sampled. Stereological techniques are then typically used to randomly select a number of sites (i.e., sampling frames) within each section, from which to obtain measurements for the several biomarkers of interest. Due to experimental resource feasibility and tissue availability considerations, only varying subsets of the 86 subjects with schizophrenia are used in individual studies.

Each individual study undergoes extensive statistical analysis. A typical approach to analyze the biomarkers under consideration has been via ANCOVA models or their multivariate version (MANCOVA). The main goal of these studies is to identify which individual biomarkers are significantly altered in subjects with schizophrenia compared with control subjects, while accounting for the pairing and the important demographic characteristics. In each study, every schizophrenia subject is matched with a control subject based upon specific demographic and other traits, namely age at death, gender, and PMI. The tissue samples obtained from a matched pair are then blinded and processed together in the tissue processing necessary in a particular study, possibly in batches of pairs.

2.2 INCORPORATING COVARIATES

In addition to variables on which control and schizophrenia subjects are paired, additional covariates are measured for each subject, such as brain tissue storage time, i.e., the amount of time that brain tissue has been stored in the Brain Tissue Bank. In analyzing individual tissue studies, the primary ANCOVA or MANCOVA models employed in CCNMD studies usually have diagnostic group as a main effect, pair as a blocking factor, and covariates such as tissue storage time. These models are considered primary due to the fact that including pair as a blocking factor usually reduces the experimental variability that arises due to the way tissue is processed. To check the robustness of the primary model, secondary ANCOVA or MANCOVA models are also typically used, in which the blocking factor of pair is replaced by the covariates on which subjects are paired, namely age at death, gender, and PMI.

2.3 MOTIVATING DATA

Initially, we were interested in ascertaining which biomarkers differ between individuals with schizophrenia and normal controls in a series of four human post-mortem tissue studies conducted by Sweet et al. [33][34][35][36], which examined in totality six biomarkers. In each of these studies, post-mortem human brain tissue, taken from the primary auditory cortex, was collected from control and schizophrenia subjects pair matched on gender and as closely as possible on age at death and PMI. Brain tissue storage time, which was not used in the matching, was also included as a covariate. Once the tissue for each subject was processed, a particular set of biomarkers was measured in multiple sections from this tissue for each study. The primary and secondary MANCOVA models described above were used in the individual studies to examine whether or not each biomarker of interest for subjects with schizophrenia differed from that of normal controls, while controlling for the effects of subject pairing and brain tissue storage time. A closer examination of this initial goal was the motivation for our research to develop a new method which could integrate data from these four studies to identify which of the six biomarkers best discriminated between the control and schizophrenia diagnostic groups, while taking into account the effects of subject

pairing and other relevant covariates on these biomarkers.

We note that there are methods that incorporate paired study design and covariates when combining results from multiple post-mortem tissue studies, as developed by Wang et al. [39]. However, these methods are not focused on group discrimination. The methodology we develop is geared towards adjusting for pairing and covariate effects when integrating paired post-mortem biomarker data, in order to better identify the biomarkers that best distinguish schizophrenia subjects from normal controls.

2.4 KONOPASKE DATA

To get a better sense of the implications of our adjustment methodology beyond the paired case, we also considered a monkey post-mortem brain tissue study conducted by Konopaske et al. [17], where we examined in our analyses six different biomarkers. In this study, brain tissue was collected from 18 male macaque monkeys that were matched in triads by terminal body weight, i.e., body weight upon sacrifice. In each triad, each monkey had been treated with one of three different drugs, sham, haloperidol, and olanzapine, the latter two of which are antipsychotics used in the treatment of schizophrenia. However, unlike the Sweet et al. data, no additional covariates were measured for these subjects. An ANOVA model was used to determine which of the biomarkers under consideration differed among the three drug groups, while controlling for the effect of group matching. In the Konopaske et al. study, there were no significant differences among the treatment groups for the noted biomarkers, other than a difference in astrocyte number between sham and antipsychotic treated subjects. Nonetheless, to illustrate our matched adjustment methodology, we apply it to the Konopaske et al. data as if to identify which of the six biomarkers best discriminate among the three drug groups of interest while, at the same time, accounting for the effect of triad matching.

3.0 ACCOUNTING FOR MATCHING AND COVARIATE EFFECTS IN LDA

3.1 CLASSIFICATION OVERVIEW

3.1.1 Bayes Classification Rule

Let $\mathbf{Y} = (Y_1, \ldots, Y_P)'$ denote a P dimensional random continuous feature vector, $\mathbf{y}$ its observed value for a particular individual, and $\mathcal{Y}$ the feature space or support of $\mathbf{Y}$. Suppose it is known that each examined individual belongs to one of g ($g \geq 2$) groups. In the context of post-mortem tissue studies, $\mathbf{Y}$ is the random vector of biomarker measurements and $\mathbf{y}$ is the observed biomarker data for an individual who belongs to one of several diagnostic groups, e.g., control or schizophrenia. The main purpose of classification is to find a rule or function of $\mathbf{y}$, which we denote as $r(\mathbf{y})$, that accurately assigns an individual with feature measurement $\mathbf{y}$ to one of these g groups. In other words, we wish to obtain a function $r(\mathbf{y})$ that optimally divides the feature space into g mutually exclusive and exhaustive regions $R_1, \ldots, R_g$ such that an individual with feature vector $\mathbf{y}$ is assigned to group i if $\mathbf{y}$ falls in R_i [7][23]. While we present our overview from the point of view of classification, our major focus is discrimination among the g groups, where our goal is to identify the most discriminatory subset of feature variables among the feature variables under consideration.

First, we denote the prior probability that an individual belongs to group i as π_i ($i = 1, \ldots, g$) and the group conditional density of $\mathbf{Y}$ in the i^{th} group as $f_i(\mathbf{y})$. Let c_{ij} be the cost of inaccurately assigning a group i individual into group j. If an individual is assigned correctly, then $c_{ii} = 0$, i.e., there is zero cost for correct assignment.

With the assumption that $\pi_1, \ldots, \pi_g$ are known and fixed, an optimal or Bayes rule $r_0(\mathbf{y})$ is a rule that has the smallest expected loss or risk among all rules $r(\mathbf{y})$ for a given $\mathbf{y}$ [1][23].

If R_i $(i = 1, \ldots, g)$ denotes the classification regions resulting from the Bayes rule $r_0(\mathbf{y})$, then [1][23]

$$R_i : \sum_{\substack{h=1 \\ h \neq i}}^{g} \pi_h f_h(\mathbf{y}) c_{hi} < \sum_{\substack{h=1 \\ h \neq j}}^{g} \pi_h f_h(\mathbf{y}) c_{hj}, \quad j = 1, \ldots, g; j \neq i. \tag{3.1}$$

In other words, $\mathbf{y}$ is assigned to the group i for which $\sum_{h=1, h\neq i}^{g} \pi_h f_h(\mathbf{y}) c_{hi}$ is minimized. If $\sum_{h=1, h\neq i}^{g} \pi_h f_h(\mathbf{y}) c_{hi}$ is minimized for two or more groups, then $\mathbf{y}$ is arbitrarily assigned to any of these groups. We note that this Bayes rule is unique if the probability of equality between the left and right hand sides of (3.1) is zero for each i and j (for each h) [1]. In the special case that the costs of misclassification c_{ij} $(i \neq j)$ are all equal, the rule in (3.1) reduces to

$$R_i : \pi_i f_i(\mathbf{y}) > \pi_j f_j(\mathbf{y}), \quad j = 1, \ldots, g; j \neq i. \tag{3.2}$$

In this case, $\mathbf{y}$ is assigned to the group i for which $\pi_i f_i(\mathbf{y})$ is maximized. If $\pi_i f_i(\mathbf{y})$ is maximized for two or more groups, then $\mathbf{y}$ is arbitrarily assigned to any of these groups. The classification regions in (3.2) can also be expressed as

$$R_i : \frac{f_i(\mathbf{y})}{f_j(\mathbf{y})} > \frac{\pi_j}{\pi_i}, \quad j = 1, \ldots, g; j \neq i. \tag{3.3}$$

In the absence of the prior probabilities π_i, Anderson[1] and McLachlan[23] discuss the conditions under which a rule can still be considered admissible, i.e., a rule that minimizes the risk attributed to the classification function $r(\mathbf{y})$ for a given $\mathbf{y}$.

Techniques that use the Bayes rule in (3.1) to determine the optimal classification regions R_i include logistic discriminant analysis [14][37], quadratic discriminant analysis (QDA) [1][23], and linear discriminant analysis, which is a special case of QDA. Our focus is on linear discriminant analysis for two or more groups, under the assumption that misclassification costs are equal.

3.1.2 Traditional Linear Discriminant Analysis for Two Groups

Let $\mathbf{Y}$ have known prior probability π_i of belonging to group i $(i = 1, \ldots, g)$, in which $\mathbf{Y} \sim N_P(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\Sigma}_{YY})$. Here, $\boldsymbol{\mu}_{Y,i} = (\mu_{Y,1,i}, \ldots, \mu_{Y,P,i})'$ is the vector of expected values for the

i^{th} group and

$$\boldsymbol{\Sigma}_{YY} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1P} \\ \vdots & \vdots & \vdots \\ \sigma_{P1} & \cdots & \sigma_{PP} \end{bmatrix}$$

is the common variance-covariance matrix in each of the g groups. The assumption of a common group variance-covariance matrix is essential to being able to write the form for the rule given in (3.3) in a simple fashion. Although the rule in (3.3) can easily be expressed in terms of our assumed distribution of $\mathbf{Y}$ in the case of multiple groups, it is natural to discuss the case of two groups due to the simplicity of the form for the rule yielded when $g = 2$ and, thus, it is this case that we discuss in detail for the rest of this section. The multiple group case is considered later in this chapter.

By taking the logarithm of both sides of (3.3) and using the monotonicity of the logarithmic function, the rule given in (3.3) can be written as follows in the case of two groups, based on the densities of $\mathbf{Y}$ in the 1^{st} and 2^{nd} groups:

$$\begin{aligned} R_1 : \left[\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_{Y,1} + \boldsymbol{\mu}_{Y,2})\right]' \boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) \geq \log(\frac{\pi_2}{\pi_1}), \\ R_2 : \left[\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_{Y,1} + \boldsymbol{\mu}_{Y,2})\right]' \boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) < \log(\frac{\pi_2}{\pi_1}), \end{aligned} \tag{3.4}$$

where $\mathbf{y}'\boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})$ is called the population linear discriminant function (LDF) [1][14][23]. If $\left[\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_{Y,1} + \boldsymbol{\mu}_{Y,2})\right]' \boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) = \log(\frac{\pi_2}{\pi_1})$, $\mathbf{y}$ could be assigned to either of the two groups; we have arbitrarily assigned $\mathbf{y}$ to group 1 in this case. In the special case that $\pi_1 = \pi_2 = 0.5$, $\log(\frac{\pi_2}{\pi_1}) = 0$.

The probability of misclassification associated with the rule in (3.4) is equal to

$$P(\mathbf{Y} \in R_1, \mathbf{Y} \in \text{group } 2) + P(\mathbf{Y} \in R_2, \mathbf{Y} \in \text{group } 1) = \pi_2 P^{(2)}(\mathbf{Y} \in R_1) + \pi_1 P^{(1)}(\mathbf{Y} \in R_2), \tag{3.5}$$

where $P^{(i)}(\mathbf{Y} \in R) = P(Y \in R | Y \in \text{group } i)$. Based on (3.5), one can easily compute the probability of misclassification, assuming $\boldsymbol{\mu}_{Y,i}$ $(i = 1, 2)$ and $\boldsymbol{\Sigma}_{YY}$ are known. Let $\mathbf{C} = \left[\mathbf{Y} - \frac{1}{2}(\boldsymbol{\mu}_{Y,1} + \boldsymbol{\mu}_{Y,2})\right]' \boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})$. We then have that $\mathbf{C} \sim N(\frac{1}{2}\boldsymbol{\Delta}^2, \boldsymbol{\Delta}^2)$ if $\mathbf{Y}$ belongs to group 1 and $\mathbf{C} \sim N(-\frac{1}{2}\boldsymbol{\Delta}^2, \boldsymbol{\Delta}^2)$ if $\mathbf{Y}$ belongs to group 2, where $\boldsymbol{\Delta}^2 = (\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})'\boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} -$

$\boldsymbol{\mu}_{Y,2}$), which is the Mahalanobis squared distance between $N_P(\boldsymbol{\mu}_{Y,1}, \boldsymbol{\Sigma}_{YY})$ and $N_P(\boldsymbol{\mu}_{Y,2}, \boldsymbol{\Sigma}_{YY})$ [1]. Based on the densities of $\mathbf{C}$ in each group, it can be shown that

$$\pi_2 P^{(2)}(\mathbf{Y} \in R_1) + \pi_1 P^{(1)}(\mathbf{Y} \in R_2) = \pi_2 \Phi\left(\frac{-\log(\frac{\pi_2}{\pi_1}) - \frac{\boldsymbol{\Delta}^2}{2}}{\boldsymbol{\Delta}}\right) + \pi_1 \Phi\left(\frac{\log(\frac{\pi_2}{\pi_1}) - \frac{\boldsymbol{\Delta}^2}{2}}{\boldsymbol{\Delta}}\right), \quad (3.6)$$

where $\Phi(\centerdot)$ is the cumulative distribution function for a standard normal random variable. If $\pi_1 = \pi_2 = 0.5$, then the formula in (3.6) reduces to $\Phi(-\frac{\boldsymbol{\Delta}}{2})$.

If the prior probabilities π_i, $\boldsymbol{\mu}_{Y,i}$, and $\boldsymbol{\Sigma}_{YY}$ are unknown, they must be estimated from sample data obtained from each of the two groups, i.e., training data. Let $\mathbf{y}_{ij} = (y_{ij,1}, \ldots, y_{ij,P})'$ be the observed feature vector for the j^{th} individual randomly sampled from the i^{th} group ($i = 1, 2$; $j = 1, \ldots, n_i$). With regards to the π_i, they may be specified in advance or, if appropriate, estimated from the training data. The sample-based counterpart of (3.4) can then be obtained by plugging in maximum likelihood (ML) estimates of $\boldsymbol{\mu}_{Y,i}$ and $\boldsymbol{\Sigma}_{YY}$, which are given by $\bar{\mathbf{y}}_i = \frac{\sum_{j=1}^{n_i} \mathbf{y}_{ij}}{n_i}$ and $\hat{\boldsymbol{\Sigma}}_{YY} = \frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)'\right]$, respectively. In addition, the unbiased estimate $\hat{\boldsymbol{\Sigma}}^*_{YY} = \frac{n_1+n_2}{n_1+n_2-2}\hat{\boldsymbol{\Sigma}}_{YY}$ can be used. The resulting sample-based rule, which can be computed using standard software packages such as SAS, is Bayes risk consistent in that its risk converges in probability, under reasonable conditions, to that of the rule given in (3.4) [37].

One way to estimate the probability of misclassification based on the rule in (3.4) is to obtain an estimate of $\boldsymbol{\Delta}$ from the training data and then plug it into the formula given in (3.6) for the probability of misclassification. Another estimation method is the resubstitution method, which involves computing the sample-based counterpart of the rule in (3.4) based on the training data and using the resulting estimated rule to predict the group membership for each individual in the training data. The proportion of individuals in the training data that are misclassified using this procedure is the resubstituted estimate of the probability of misclassification associated with (3.4). However, this estimate is asymptotically biased due to the fact that it is computed using the same sample that was used to construct the sample LDF in the first place [22][23]. To considerably reduce this bias, we can instead use V-fold cross validation, where V ranges from 2 to the total sample size [12][23]. In this procedure, the training data are first randomly split into V mutually exclusive subsets of approximately equal size, where each of the V subsets are then dropped out while the

remaining $V-1$ subsets are used to compute the estimates of $\boldsymbol{\mu}_{Y,i}$ and $\boldsymbol{\Sigma}_{YY}$. Once these estimates are plugged into (3.4), the resulting estimated rule is used to predict the group membership for each individual in the omitted subset. The cross validated estimate of the probability of misclassification is then computed as $\pi_1 p_{21} + \pi_2 p_{12}$, where p_{ij} is the proportion of group j individuals in the training data that are misclassified into group i in this manner $(i, j = 1, 2; i \neq j)$. Both the resubstitution and cross validation estimation methods can be easily implemented using standard software packages.

The raw discriminant coefficients, $\boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})$, are highly dependent on the measurement scale of the feature data. As a result, it is often desirable to standardize these coefficients in order to accurately determine which feature variables have high classifying significance relative to the others in the linear discriminant function. Each discriminant coefficient is standardized by taking the product of its original value and the feature variable's standard deviation. The feature variables whose standardized discriminant coefficients are fairly large in absolute value are those that best discriminate between the first and second groups. Similarly, for unknown $\boldsymbol{\mu}_{Y,i}$ and $\boldsymbol{\Sigma}_{YY}$, it is often preferable to standardize the estimated raw discriminant coefficients, $\hat{\boldsymbol{\Sigma}}_{YY}^{-1}(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)$. Each estimated discriminant coefficient is standardized by taking the product of its original value and the feature variable's estimated standard deviation, pooled across the groups under consideration [26][27]. We note that these standardized discriminant coefficients are also equal to the estimated raw discriminant coefficients obtained from implementing traditional LDA on the standardized training data $\mathbf{y}_{ij}^{std} = (y_{ij,1}^{std}, \ldots, y_{ij,P}^{std})'$. If $\hat{\sigma}_{pp}$ denotes the estimated pooled variance for the p^{th} feature variable, then $y_{ij,p}^{std} = \frac{1}{\sqrt{\hat{\sigma}_{pp}}}(y_{ij,p} - \bar{y}_{..,p})$, where $\bar{y}_{..,p}$ is the sample mean for the p^{th} feature variable. In this case, the sample-based counterpart of (3.4) is given by

$$R_1 : \hat{\mathbf{L}}'\mathbf{y}^{\mathbf{std}} \geq \log(\frac{\pi_2}{\pi_1}), \quad R_2 : \hat{\mathbf{L}}'\mathbf{y}^{\mathbf{std}} < \log(\frac{\pi_2}{\pi_1}),$$

where $\mathbf{y}^{\mathbf{std}} = (y_1^{std}, \ldots, y_P^{std})'$ and $\hat{\mathbf{L}}$ is the P dimensional vector of estimated standardized discriminant coefficients. The sign of the discriminant coefficient for the p^{th} feature variable $(p = 1, \ldots, P)$ can then be used to determine whether relatively large or small values are associated with group 1 compared with group 2, holding all other feature variables fixed.

3.2 COVARIANCE ADJUSTED LINEAR DISCRIMINANT ANALYSIS

3.2.1 Motivation

In addition to determining a subset of feature variables that best discriminates among several groups of interest, we seek to be able to classify an individual into one of these groups. We do note that in the application to post-mortem tissue studies, the focus is solely on discrimination, not classification.

However, the distribution of the feature data $\mathbf{Y}$ may depend on a particular set of covariates $\mathbf{X} = (X_1, \ldots, X_S)'$. Therefore, it is necessary to control or adjust for these covariate effects in order to accurately determine the true discriminatory power of the feature data. To illustrate this fact, we present the following scenario in Figure 3.1 for univariate $\mathbf{Y}$ and $\mathbf{X}$. Here, we wish to discriminate between control and schizophrenia subjects, where each subject has observed biomarker measurement y and storage time value x.

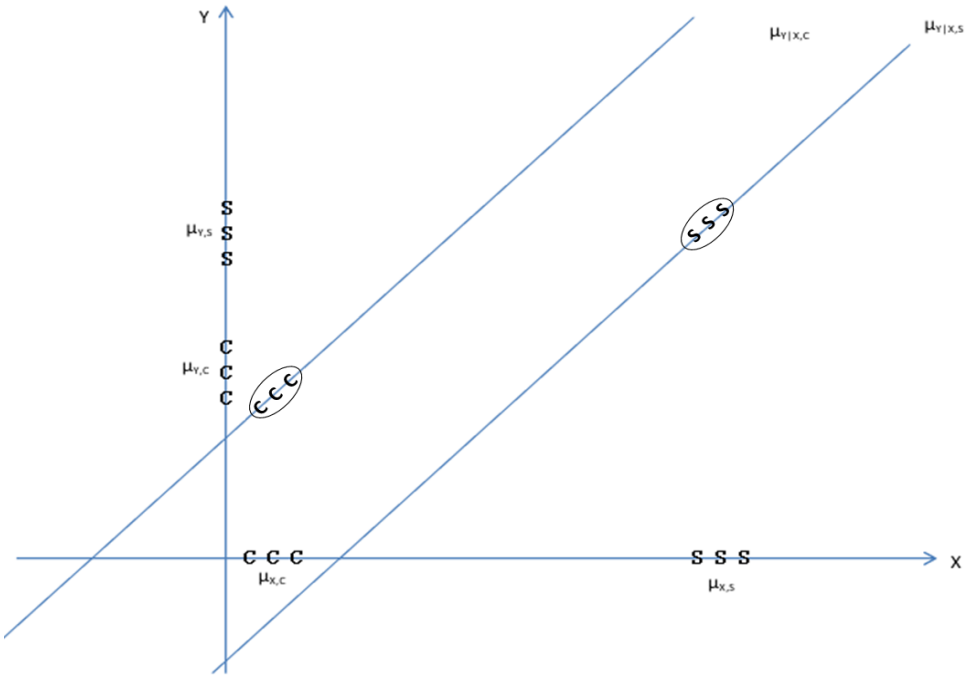

Figure 3.1: Plot of y vs. x, along with conditional means $\mu_{y|x,c}$ and $\mu_{y|x,s}$

Based on the marginal distribution of $\mathbf{Y}$, we see that, on average, large y values correspond to the schizophrenia group while small y values correspond to the control group. Thus, new subjects will be classified into the schizophrenia group if they have relatively large y values and into the control group otherwise. If we examine the joint distribution of $\mathbf{Y}$ and

$\mathbf{X}$, we see that, on average, large values for y and x correspond to the schizophrenia group while small values correspond to the control group. New subjects will then be classified into the schizophrenia group if they have relatively large y and x values and into the control group otherwise. Storage time can be viewed as being experimentally controlled and, thus, is also extraneous to the clinical issue of interest as to whether there is a difference in the biomarker $\mathbf{Y}$ between subjects with schizophrenia and controls. However, we do see that the distribution of $\mathbf{X}$ depends on group in this case. Therefore, the discriminatory power of $\mathbf{Y}$ may be clouded by the group differences in $\mathbf{X}$. However, if we examine the distribution of $\mathbf{Y}$ given $\mathbf{X}$, we see that the conditional mean of $\mathbf{Y}$ is higher in the control group. Thus, if storage time is fixed at a particular value, new subjects will be classified into the control group if they have relatively large y values and into the schizophrenia group otherwise.

From this scenario, we see that in the presence of covariate effects, the only way to get a clear picture of how well $\mathbf{Y}$ discriminates between the two diagnostic groups is to focus on the conditional distributions of $\mathbf{Y}$ given $\mathbf{X}$ for each group. Cochran and Bliss[8], Lachenbruch[19], and Tu et al. [37] recognized this fact and developed covariance adjusted linear discriminant analysis to account for covariate effects. In Sections 3.2.2 and 3.2.3, we summarize their methods, which use the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ to determine the discriminatory power of the feature data without the confounding effects of the covariates one is not primarily interested in. Although we only consider the case of two groups for the sake of notational simplicity, we note that covariance adjusted LDA can be easily extended to handle more than two groups, an extension we develop in greater detail in Sections 3.5 and 3.6 for matching across multiple groups.

3.2.1.1 Bayes Conditional Classification Rule Let $f_i(\mathbf{y}|\mathbf{x})$ denote the conditional density of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ in the i^{th} group ($i = 1, \ldots, g$). If we assume equal misclassification costs, then $f_i(\mathbf{y})$ can be replaced with $f_i(\mathbf{y}|\mathbf{x})$ in (3.3) to give the following rule, on which the development of covariance adjusted LDA is based, that is used to classify an individual whose covariate vector $\mathbf{x}$ has been observed in correspondence with feature vector $\mathbf{y}$ [19][23]:

$$R_i : \frac{f_i(\mathbf{y}|\mathbf{x})}{f_j(\mathbf{y}|\mathbf{x})} > \frac{\pi_j}{\pi_i}, \quad j = 1, \ldots, g; j \neq i. \tag{3.7}$$

3.2.2 Traditional Covariance Adjusted LDA for Two Groups

Cochran and Bliss[8] assume that for the i^{th} group ($i = 1, 2$), $(\mathbf{Y}, \mathbf{X}) \sim N_{P\times S}(\boldsymbol{\mu}_{Y,X,i}, \boldsymbol{\Sigma}_{Y,X})$, where:

$$\boldsymbol{\mu}_{Y,X,i} = \left(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\mu}_X\right) \quad \text{and} \quad \boldsymbol{\Sigma}_{Y,X} = \begin{pmatrix} \boldsymbol{\Sigma}_{YY} & \boldsymbol{\Sigma}_{YX} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{XX} \end{pmatrix}.$$

Thus, the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ in the i^{th} group is multivariate normal with mean vector

$$\boldsymbol{\mu}_{Y|X,i} = \boldsymbol{\mu}_{Y,i} + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\mathbf{x} - \boldsymbol{\mu}_X) = \boldsymbol{\tau}_{Y,i} + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\mathbf{x} \ (\tau_{Y,i} = \boldsymbol{\mu}_{Y,i} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\mu}_X)$$

and variance-covariance matrix $\boldsymbol{\Sigma}_{Y|X} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}$ that is common to both groups.

If we assume equal misclassification costs, then the rule given in (3.7) can be written as follows, based on the conditional densities of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$:

$$\begin{aligned} R_1: \ & \left[\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_{Y|X,1} + \boldsymbol{\mu}_{Y|X,2})\right]' \boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y|X,1} - \boldsymbol{\mu}_{Y|X,2}) \geq \log(\frac{\pi_2}{\pi_1}), \\ R_2: \ & \left[\mathbf{y} - \frac{1}{2}(\boldsymbol{\mu}_{Y|X,1} + \boldsymbol{\mu}_{Y|X,2})\right]' \boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y|X,1} - \boldsymbol{\mu}_{Y|X,2}) < \log(\frac{\pi_2}{\pi_1}), \end{aligned} \tag{3.8}$$

which can also be expressed as

$$\begin{aligned} R_1: \ & \left[\tilde{\mathbf{y}} - \frac{1}{2}(\boldsymbol{\mu}_{Y,1} + \boldsymbol{\mu}_{Y,2})\right]' \boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) \geq \log(\frac{\pi_2}{\pi_1}), \\ R_2: \ & \left[\tilde{\mathbf{y}} - \frac{1}{2}(\boldsymbol{\mu}_{Y,1} + \boldsymbol{\mu}_{Y,2})\right]' \boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) < \log(\frac{\pi_2}{\pi_1}), \end{aligned} \tag{3.9}$$

using the formula for $\boldsymbol{\mu}_{Y|X,i}$, where $\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\mathbf{x} - \boldsymbol{\mu}_X)$. Conditional on $\mathbf{x}$, $\tilde{\mathbf{Y}} = \mathbf{Y} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}(\mathbf{x} - \boldsymbol{\mu}_X) \sim N_P(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\Sigma}_{Y|X})$ in the i^{th} group, which implies that the rule based on the densities of $\tilde{\mathbf{Y}}$ can be expressed in the same form as that in (3.4) obtained from traditional LDA, where the observed $\mathbf{y}$ is now suitably adjusted for all covariate effects and the conditional variance-covariance matrix is used.

From (3.9), we obtain the vector of adjusted discriminant coefficients $\boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})$, which we can use to identify the feature variables in $\mathbf{Y}$ that best discriminate between the two groups, while accounting for all relevant covariate effects.

Suppose π_i, $\boldsymbol{\mu}_{Y,X,i}$, and $\boldsymbol{\Sigma}_{Y,X}$ are unknown. We then have that $\boldsymbol{\Sigma}_{Y|X}$ is unknown and that $\boldsymbol{\mu}_{Y|X,i}$ is an unknown function of $\mathbf{x}$. In this case, we must use training data consisting of $(\mathbf{y}_{ij}, \mathbf{x}_{ij})$, the observed feature and covariate vectors for the j^{th} individual sampled from the i^{th} group ($i = 1, 2$; $j = 1, \ldots, n_i$), to estimate $\boldsymbol{\mu}_{Y|X,i}$ and $\boldsymbol{\Sigma}_{Y|X}$. From the training data, we can obtain the ML estimates $\hat{\boldsymbol{\mu}}_{Y,i}$, $\hat{\boldsymbol{\mu}}_X$, $\hat{\boldsymbol{\Sigma}}_{YY}$, $\hat{\boldsymbol{\Sigma}}_{YX}$, and $\hat{\boldsymbol{\Sigma}}_{XX}$. The sample-based counterparts of (3.8) and (3.9) can then be obtained by plugging in the ML estimates $\hat{\boldsymbol{\mu}}_{Y|X,i}$ $= \hat{\boldsymbol{\mu}}_{Y,i} + \hat{\boldsymbol{\Sigma}}_{YX}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}_X)$ and $\hat{\boldsymbol{\Sigma}}_{Y|X} = \hat{\boldsymbol{\Sigma}}_{YY} - \hat{\boldsymbol{\Sigma}}_{YX}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}\hat{\boldsymbol{\Sigma}}_{XY}$ [23]. The prior probabilities π_i may be obtained in the same manner as described in Section 3.1.2.

It is easy to show that the probability of misclassification based on the conditional distributions of $\mathbf{Y}$ for a given $\mathbf{x}$ is equal to $\pi_2\Phi\left(\frac{-\log(\frac{\pi_2}{\pi_1}) - \frac{\boldsymbol{\alpha}^2}{2}}{\boldsymbol{\alpha}}\right) + \pi_1\Phi\left(\frac{\log(\frac{\pi_2}{\pi_1}) - \frac{\boldsymbol{\alpha}^2}{2}}{\boldsymbol{\alpha}}\right)$, where $\boldsymbol{\alpha}^2 = (\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})'\boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})$. In particular, in the case where $\pi_1 = \pi_2 = 0.5$, this probability of misclassification reduces to $\Phi(-\frac{\boldsymbol{\alpha}}{2})$. Assuming $\mathbf{Y}$ and $\mathbf{X}$ are not independent, it can be directly shown that $\boldsymbol{\Sigma}_{Y|X}^{-1} - \boldsymbol{\Sigma}_{YY}^{-1}$ is positive definite and, thus,

$$\begin{aligned}
\boldsymbol{\alpha}^2 &= (\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})'\boldsymbol{\Sigma}_{Y|X}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) \\
&> (\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2})'\boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{\mu}_{Y,1} - \boldsymbol{\mu}_{Y,2}) \\
&= \boldsymbol{\Delta}^2.
\end{aligned}$$

It then follows that $\Phi(-\frac{\alpha}{2}) < \Phi(-\frac{\boldsymbol{\Delta}}{2})$ in this case. Therefore, for equal priors, we have that if the distribution of $\mathbf{X}$ and the relationship between $\mathbf{Y}$ and $\mathbf{X}$ are not group dependent, then conditioning on $\mathbf{X}$ produces lower population misclassification rates compared to those obtained when $\mathbf{X}$ is ignored, a result first noted by Cochran and Bliss [8][19][37].

We note that traditional covariance adjusted LDA, as formulated by Cochran and Bliss, is actually a special case of traditional covariance adjusted QDA, which is fully described by Rawlings et al. in their discussion of conditional quadratic discrimination [29].

3.2.3 General Covariance Adjusted LDA for Two Groups

It is not always the case that the conditional mean of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is a linear function of $\mathbf{x}$ or that $\mathbf{Y}$ and $\mathbf{X}$ are jointly multivariate normal. Lachenbruch [19] and Tu et al. [37] relax the assumption of joint normality of $\mathbf{Y}$ and $\mathbf{X}$ and, instead, only assume that given $\mathbf{X} = \mathbf{x}$,

$\mathbf{Y} \sim N_P(h_i(\mathbf{x}), \boldsymbol{\Sigma})$ in the i^{th} group $(i = 1, 2)$, where $\boldsymbol{\Sigma}$ is the common conditional variance-covariance matrix in each of the two groups, $h_i(\mathbf{x}) = \boldsymbol{\mu}_i + \boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$, $\boldsymbol{\mu}_i = (\mu_{1,i}, \ldots, \mu_{P,i})'$ corresponds to the effect of the i^{th} group on $\mathbf{Y}$, and $\boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta}) = (\rho_1(\mathbf{x}; \boldsymbol{\theta}_1), \ldots, \rho_P(\mathbf{x}; \boldsymbol{\theta}_P))'$ is a known smooth function of a given $\mathbf{x}$ that does not depend on group, but does depend on parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ corresponding, respectively, to each of the P feature variables in $\mathbf{Y}$. We have that $\boldsymbol{\mu}_i \in \mathbb{R}^P$, $\boldsymbol{\theta}_p \in \mathbb{R}^{D_p}$ $(D_p \in \mathbb{N};\ p = 1, \ldots, P)$, and $\boldsymbol{\Sigma}$ is assumed to be positive definite. The assumption regarding the conditional distribution of $\mathbf{Y}$ made by Lachenbruch and Tu et al. is a generalization of the assumption made by Cochran and Bliss in Section 3.2.2, namely, $\mathbf{Y}|\mathbf{x} \sim N_P(\boldsymbol{\mu}_{Y|X,i}, \boldsymbol{\Sigma}_{Y|X})$ in the i^{th} group, where the conditional mean is given by $\boldsymbol{\mu}_{Y|X,i} = \boldsymbol{\tau}_{Y,i} + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\mathbf{x}$ $(\tau_{Y,i} = \boldsymbol{\mu}_{Y,i} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\mu}_X)$. Although we do not provide the details in this discussion, we point out that general covariance adjusted LDA can be extended to handle QDA by using Lachenbruch and Tu et al.'s approach to generalize the assumptions made by Rawlings et al. in their development of traditional covariance adjusted QDA [37].

Based on the conditional densities of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$, the rule given in (3.7) can be written as follows:

$$\begin{aligned} R_1 :\ & \left\{\mathbf{y} - \frac{1}{2}\left[h_1(\mathbf{x}) + h_2(\mathbf{x})\right]\right\}' \boldsymbol{\Sigma}^{-1}\left[h_1(\mathbf{x}) - h_2(\mathbf{x})\right] \geq \log(\frac{\pi_2}{\pi_1}), \\ R_2 :\ & \left\{\mathbf{y} - \frac{1}{2}\left[h_1(\mathbf{x}) + h_2(\mathbf{x})\right]\right\}' \boldsymbol{\Sigma}^{-1}\left[h_1(\mathbf{x}) - h_2(\mathbf{x})\right] < \log(\frac{\pi_2}{\pi_1}). \end{aligned} \tag{3.10}$$

Using the formulas for $h_i(\mathbf{x})$, we have that $\frac{1}{2}\left[h_1(\mathbf{x}) + h_2(\mathbf{x})\right] = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$ and $h_1(\mathbf{x}) - h_2(\mathbf{x}) = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Thus, the classification regions in (3.10) can be re-expressed as:

$$\begin{aligned} R_1 :\ & \left[\tilde{\mathbf{y}} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \log(\frac{\pi_2}{\pi_1}), \\ R_2 :\ & \left[\tilde{\mathbf{y}} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \log(\frac{\pi_2}{\pi_1}), \end{aligned} \tag{3.11}$$

where $\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$. Given $\mathbf{X} = \mathbf{x}$, $\tilde{\mathbf{Y}} = \mathbf{Y} - \boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta}) \sim N_P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ in the i^{th} group. As was the case for traditional covariance adjusted LDA, we see from (3.11) that the classification rule based on the densities of $\tilde{\mathbf{Y}}$ can be expressed in the same form as that in (3.4) for traditional LDA, where the observed $\mathbf{y}$ is adjusted for all covariate effects and the conditional variance-covariance matrix is used.

Once they have been suitably standardized, the elements of the adjusted discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ can be used to determine the feature variables that best discriminate between the first and second groups, taking into account the effects of the covariate vector $\mathbf{X}$.

According to Tu et al. [37], the sample-based counterparts of (3.10) and (3.11) can be obtained by plugging in consistent estimators of $h_i(\mathbf{x})$ and $\boldsymbol{\Sigma}$. A consistent estimator of $h_i(\mathbf{x})$ is given by $\hat{h}_i(\mathbf{x}) = \hat{\boldsymbol{\mu}}_i + \boldsymbol{\rho}(\mathbf{x}; \hat{\boldsymbol{\Theta}}) = (\hat{\mu}_{1,i} + \rho_1(\mathbf{x}; \hat{\boldsymbol{\theta}}_1), \ldots, \hat{\mu}_{P,i} + \rho_P(\mathbf{x}; \hat{\boldsymbol{\theta}}_P))'$, where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ are consistent estimators of $\boldsymbol{\mu}_i$ and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$, e.g., ML or least squares (LS) estimators, that are computed from the training data $(\mathbf{y}_{ij}, \mathbf{x}_{ij})$. Depending on the structure of $\rho_p(\mathbf{x}; \boldsymbol{\theta}_p)$ $(p = 1, \ldots, P)$, the parameter vectors $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ may not be identifiable, which Tu et al. do not address in their discussion. If this is the case, then the estimates $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ are not unique, which could mean that the sample-based counterpart of (3.10) may vary depending on the values of these estimates.

In certain cases, however, the sample-based counterpart of (3.10) remains invariant even if $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ are not identifiable. Specifically, if $\boldsymbol{\rho}(\centerdot; \boldsymbol{\Theta})$ is a linear function of $\boldsymbol{\Theta}$, then $h_i(\mathbf{x})$ is an estimable function of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Theta}$. Also, under sufficient regularity conditions[18], $h_{p,i}(\mathbf{x}) = \mu_{p,i} + \rho_p(\mathbf{x}; \boldsymbol{\theta}_p)$ is an estimable function of $\boldsymbol{\mu}_{p,i}$ and θ_p if $\rho_p(\centerdot; \boldsymbol{\theta}_p)$ is a nonlinear function $(p = 1, \ldots, P)$. In either of these two instances, $h_i(\mathbf{x})$ is an estimable function of $\boldsymbol{\mu}_i$ and $\boldsymbol{\Theta}$ for a given $\mathbf{x}$, which implies that the estimates $\hat{h}_i(\mathbf{x})$ $(i = 1, 2)$, $\hat{h}_1(\mathbf{x}) + \hat{h}_2(\mathbf{x})$, and $\hat{h}_1(\mathbf{x}) - \hat{h}_2(\mathbf{x})$ are unique. Also, we later show that the estimability of $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ implies that the estimate of $\boldsymbol{\Sigma}$ is also unique. Therefore, in these instances where $h_i(\mathbf{x})$ is estimable, the sample-based counterpart of (3.10) remains invariant for a given $\mathbf{x}$ even when the estimates $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ are not unique.

If $\mathbf{x}_{ij}$, $\hat{\boldsymbol{\mu}}_i$, and $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ are held fixed, the covariate adjusted training feature data $\hat{\tilde{\mathbf{y}}}_{ij} = \mathbf{y}_{ij} - \boldsymbol{\rho}(\mathbf{x}_{ij}; \hat{\boldsymbol{\Theta}}) = \Big(y_{ij,1} - \rho_1(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_1), \ldots, y_{ij,P} - \rho_P(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_P)\Big)'$ constitute a random sample with mean $\hat{\boldsymbol{\mu}}_i$ in the i^{th} group. Using this fact, but with little attention to estimability issues, Tu et al. [37] argue that conditional on $\mathbf{x}_{ij}$, $\hat{\boldsymbol{\mu}}_i$, and $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$, a consistent estimator of $\boldsymbol{\Sigma}$

is given by

$$\begin{aligned}\hat{\boldsymbol{\Sigma}} &= \frac{1}{n_1+n_2-2}\left[\sum_{j=1}^{n_1}(\mathbf{y}_{1j}-\hat{h}_1(\mathbf{x}_{1j}))(\mathbf{y}_{1j}-\hat{h}_1(\mathbf{x}_{1j}))' + \sum_{j=1}^{n_2}(\mathbf{y}_{2j}-\hat{h}_2(\mathbf{x}_{2j}))(\mathbf{y}_{2j}-\hat{h}_2(\mathbf{x}_{2j}))'\right] \\ &= \frac{1}{n_1+n_2-2}\left[\sum_{j=1}^{n_1}(\hat{\tilde{\mathbf{y}}}_{1j}-\hat{\boldsymbol{\mu}}_1)(\hat{\tilde{\mathbf{y}}}_{1j}-\hat{\boldsymbol{\mu}}_1)' + \sum_{j=1}^{n_2}(\hat{\tilde{\mathbf{y}}}_{2j}-\hat{\boldsymbol{\mu}}_2)(\hat{\tilde{\mathbf{y}}}_{2j}-\hat{\boldsymbol{\mu}}_2)'\right],\end{aligned} \tag{3.12}$$

which we observe is unique if $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ are estimable. Once we plug in the estimates $\hat{h}_i(\mathbf{x})$ and $\hat{\boldsymbol{\Sigma}}$, the sample counterpart of (3.10) can be written as

$$\begin{aligned} R_1 &: \left[\hat{\tilde{\mathbf{y}}} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)\right]' \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2) \geq \log(\frac{\pi_2}{\pi_1}), \\ R_2 &: \left[\hat{\tilde{\mathbf{y}}} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)\right]' \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1-\hat{\boldsymbol{\mu}}_2) < \log(\frac{\pi_2}{\pi_1}), \end{aligned} \tag{3.13}$$

where $\hat{\tilde{\mathbf{y}}} = \mathbf{y} - \boldsymbol{\rho}(\mathbf{x};\hat{\boldsymbol{\Theta}})$.

Tu et al. discuss how the resubstituted estimate of the probability of misclassification for the rule in (3.10) can be computed in a manner similar to that used in traditional LDA. To clarify, using the training data $(\mathbf{y}_{ij}, \mathbf{x}_{ij})$, the model for the conditional mean $h_i(\mathbf{x}_{ij}) = \boldsymbol{\mu}_i + \boldsymbol{\rho}(\mathbf{x}_{ij};\boldsymbol{\Theta})$ is first fit so that we may obtain the consistent estimators $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$, which we then use to compute the covariate adjusted training feature data $\hat{\tilde{\mathbf{y}}}_{ij} = \mathbf{y}_{ij} - \boldsymbol{\rho}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}})$. From this adjusted data, we can obtain the estimates $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$ as described above and subsequently compute the estimated rule in (3.13). Using this estimated rule, we predict the group membership for each individual in the training data based on the value of $\hat{\tilde{\mathbf{y}}}_{ij}$, and compute the proportion of individuals that are misclassified, which is the resubstituted estimate of the misclassification probability for the rule in (3.10). However, because this estimate is computed using the same sample used to obtain the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$, it underestimates the true probability of misclassification [37].

As with traditional LDA, V-fold cross validation may help to reduce this bias [37]. In particular, once the covariate adjusted training feature data $\hat{\tilde{\mathbf{y}}}_{ij}$ are computed as described above, we split them into V mutually exclusive subsets of approximately equal size. Each of these V subsets are then dropped out while the remaining $V-1$ subsets are used to compute the estimates $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, and $\hat{\boldsymbol{\Sigma}}$. Based on these estimates, we can compute the estimated rule in (3.13) and use it to predict the group membership for each individual in the omitted subset

based on his or her value of $\hat{\hat{\mathbf{y}}}_{ij}$. As in traditional LDA, the cross validated estimate of the probability of misclassification for the rule in (3.10) is computed as $\pi_1 p_{21} + \pi_2 p_{12}$, where p_{ij} is the proportion of group j individuals that are misclassified into group i in this manner $(i, j = 1, 2; i \neq j)$.

3.2.4 Summary of Covariance Adjusted LDA

Both the traditional and general versions of covariance adjusted LDA allow us to account for the effects of the covariate vector $\mathbf{X}$ when using LDA methods to develop a rule from which we can determine the most discriminatory subset of feature variables and classify new individuals. With training data, this covariate adjusted classification rule can be easily computed using any available software package that implements traditional LDA, thereby eliminating the need to develop new software for covariance adjusted LDA.

The main difference between traditional covariance adjusted LDA and general covariance adjusted LDA is that the general approach makes no assumptions regarding the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$ when it controls for covariate effects, which is beneficial with regards to our research focus on post-mortem tissue studies. In these studies, biomarker measurements are taken from control and schizophrenia subjects that are paired based on specific characteristics and measured for additional covariates. When analyzing such data, we do not assume that pair membership, biomarker values, and covariate values have a joint distribution and, thus, the conditional model employed by Lachenbruch and Tu et al. is more appealing with regards to our research than the jointly normal model employed by Cochran and Bliss.

However, in their development of covariance adjusted LDA, none of these authors accounted for the fact that individuals may be matched on certain characteristics, where such matching can also have an effect on the feature variables under consideration. For example, in post-mortem brain tissue studies comparing the neurobiological characteristics of normal controls and subjects with schizophrenia, subjects are typically paired on a number of demographic variables. In these studies, the biomarkers of interest may depend not only on covariates, such as brain tissue storage time, but also on the methods used to process the brain tissue for each subject pair. Therefore, the work of these authors requires an extension to handle subject matching, which we detail in Sections 3.3 to 3.6. Due to the fact

that general covariance adjusted LDA is more applicable to our research than its traditional counterpart, we extend this formulation in Section 3.3 under the assumption that individuals are paired on certain characteristics without any additional covariates. In Section 3.4, we extend the methods developed in Section 3.3 to the case where paired individuals are also measured on additional covariates. Our methodology to account for matching and covariate effects is extended to the case of multiple groups in Sections 3.5 and 3.6.

3.3 PAIRED LINEAR DISCRIMINANT ANALYSIS

3.3.1 Normal Populations with Known Parameters

In the general case where individuals are matched across g ($g \geq 2$) different groups, we begin by considering the conditional distribution of the feature vector $\mathbf{Y}$ for a specific g-tuple or match. We introduce a parameter vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_P)'$, which corresponds to each individual in a match across the P feature variables and denotes the effect of group matching on $\mathbf{Y}$. For example, in post-mortem tissue processing, γ_p may represent the strengths of the processing solutions used in obtaining the p^{th} biomarker of interest. Recall that in Lachenbruch and Tu et al.'s model for the conditional mean, all of the model parameters could be estimated from available training data, after which these estimates can be plugged into the covariate adjusted linear discriminant rule in (3.10) to classify a new individual. In our formulation to account for matching, we include the parameter $\boldsymbol{\gamma}$ in this model, which can be estimated for each individual in each match in the training data. However, for each individual in a new match beyond the training data, where, for example, a new tissue processing solution may be used, $\boldsymbol{\gamma}$ must be re-estimated because it is specific to each match under consideration. One can then view our matched adjustment methodology as an extension of the conditional model under general covariance adjusted LDA to also include parameters that are specific to each member of a particular match and, thus, must be re-estimated for each new match.

We first present the case in which individuals are paired across two groups, where our discussion focuses on three different approaches that can be taken to account for the effect

of pairing from a population based standpoint. When given a set of feature measurements for a particular pair, we know that one member belongs to the first group while the other member belongs to the second group. In this case, it is equally likely that each pair member belongs to either of the two groups because we assume there is no preference for which pair member is designated as being the first or second member. Thus, for each pair, we assume that the feature vector for each of the two members has prior probability 0.5 of belonging to either group, i.e., $\pi_1 = \pi_2 = 0.5$. In addition, we assume equal misclassification costs.

3.3.1.1 Classifying One Pair Member using Known Pair Effect

To account for the effect of pairing, one approach we can take is to apply our previously described extension of Lachenbruch and Tu et al.'s conditional model to the random feature vector for any individual in a pair, namely, $\mathbf{Y}_{ind}$. In other words, for known $\boldsymbol{\gamma}$, we assume $\mathbf{Y}_{ind} \sim N_P(\boldsymbol{\mu}_i + \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ in the i^{th} group $(i = 1, 2)$, where $\boldsymbol{\Sigma}$ is assumed to be positive definite. We note that our application of this model based solely on $\mathbf{Y}_{ind}$ only makes sense in a population setting. In practice, we can only apply our model if we consider the feature vectors for both members of a particular pair, as we show in Section 3.3.2.1.

Based on the densities of $\mathbf{Y}_{ind}$, we can apply LDA to obtain the following classifcation regions

$$\begin{aligned} R_1 : \ & \left[\tilde{\mathbf{y}}_{ind} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \\ R_2 : \ & \left[\tilde{\mathbf{y}}_{ind} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0, \end{aligned} \tag{3.14}$$

where $\tilde{\mathbf{y}}_{ind} = \mathbf{y}_{ind} - \boldsymbol{\gamma}$ represents the feature measurement for an individual that has been adjusted for the effect of pairing. Using this classification rule, we have that an individual with the adjusted feature measurement $\tilde{\mathbf{y}}_{ind}$ is assigned to group 1 if $\tilde{\mathbf{y}}_{ind}$ falls in R_1 and to group 2 otherwise.

More importantly, once its coefficients have been suitably standardized, the linear discriminant function $\tilde{\mathbf{y}}'_{ind}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in (3.14) can be used to determine which of the feature variables under consideration best discriminate between groups 1 and 2, adjusting for the effect of pairing. The sign of the p^{th} $(p = 1, \ldots, P)$ element of the discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ can be used to determine whether the p^{th} adjusted feature variable

is larger or smaller for group 1 compared with group 2, holding all other adjusted feature variables fixed.

We show later on in Section 3.3.2.1 that with training data, which are collected in pairs, one can estimate $\boldsymbol{\gamma}$ for each training pair, along with the values of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$, and plug these estimates into (3.14) to predict the group membership of each member of each training pair. On the other hand, for an individual belonging to a pair not included in the training data, we must re-estimate $\boldsymbol{\gamma}$ in order to use the rule in (3.14) to classify this individual. However, as we show in Section 3.3.2.1, it may not be possible to estimate $\boldsymbol{\gamma}$ if we're only provided with the feature data for this one individual.

3.3.1.2 Classifying One Pair Member using Pairwise Feature Difference

A rather intuitive alternative to the previous approach we develop to account for the effect of pairing on the feature data is to implement traditional LDA on the pairwise differenced random feature vector $\mathbf{Y}_{ind} - \mathbf{Y}_{sib} \equiv \mathbf{Y}_{\text{diff}}$ for an individual in a pair, as used by Wang et al. [38], where $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ are the random feature vectors that correspond to an individual and their sibling in a pair.

Specifically, we assume $\mathbf{Y}_{\text{diff}} \sim N_P(\boldsymbol{\mu}_i^{\text{diff}}, \boldsymbol{\Sigma}_*)$ in the i^{th} population ($i = 1, 2$), where $\boldsymbol{\mu}_1^{\text{diff}} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_2^{\text{diff}} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$. If $\mathbf{Y}_{\text{diff}}$ belongs to the i^{th} population, then $\mathbf{Y}_{ind}$ belongs to the i^{th} group. Regardless of whether the variance-covariance matrices for $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ are the same in each population, $\boldsymbol{\Sigma}_*$ remains the same for each of the two populations. Also, whether or not the covariance matrix between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ is symmetric, $\boldsymbol{\Sigma}_*$ is still common to each population of $\mathbf{Y}_{\text{diff}}$.

For a given pair, each member is equally likely to belong to either of the two groups and the labeling of a member as an individual or a sibling is assumed to be completely random. Based on these two facts, and the fact that the feature vector $\mathbf{Y}_{ind}$ for an individual in a pair belongs to the i^{th} group in the i^{th} population of $\mathbf{Y}_{\text{diff}}$, we assume that the prior probability of each population is 0.5. The rule in (3.4) used for traditional LDA can then be applied to $\mathbf{Y}_{\text{diff}}$ to obtain

$$\begin{aligned} R_1^{\text{diff}}: \quad & \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \frac{1}{2}\left(\boldsymbol{\mu}_1^{\text{diff}} + \boldsymbol{\mu}_2^{\text{diff}}\right)\right]' \boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\mu}_1^{\text{diff}} - \boldsymbol{\mu}_2^{\text{diff}}\right) \geq 0, \\ R_2^{\text{diff}}: \quad & \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \frac{1}{2}\left(\boldsymbol{\mu}_1^{\text{diff}} + \boldsymbol{\mu}_2^{\text{diff}}\right)\right]' \boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\mu}_1^{\text{diff}} - \boldsymbol{\mu}_2^{\text{diff}}\right) < 0, \end{aligned} \tag{3.15}$$

which, after some simplification, reduces to

$$R_1^{\text{diff}}: \ (\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \quad R_2^{\text{diff}}: \ (\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0. \quad (3.16)$$

For each pair member that we wish to classify, we first compute the pairwise feature difference $\mathbf{y}_{ind} - \mathbf{y}_{sib}$, the difference between the feature measurement for that pair member and that of their sibling. If this difference falls into region R_1^{diff}, then this difference is assigned to the first population and, thus, we classify that pair member into the first group. Otherwise, if $\mathbf{y}_{ind} - \mathbf{y}_{sib}$ falls into region R_2^{diff}, then this difference is assigned to the second population and we classify that pair member into the second group. In classifying each individual in a pair using $\mathbf{y}_{ind} - \mathbf{y}_{sib}$, we have that their sibling is classified using the difference $\mathbf{y}_{sib} - \mathbf{y}_{ind} = -(\mathbf{y}_{ind} - \mathbf{y}_{sib})$. Based on the rule in (3.16), for which the discriminant function $(\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is compared to the cutpoint of zero, the fact that $\mathbf{y}_{sib} - \mathbf{y}_{ind} = -(\mathbf{y}_{ind} - \mathbf{y}_{sib})$ ensures that if an individual in a pair is classified into the first group, then their sibling is classified into the second group, and vice versa.

Once its coefficients have been properly standardized, the discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in (3.16) can be used to identify the feature variables that best distinguish an individual belonging to group 1 from that belonging to group 2 in any given pair. The sign of the p^{th} $(p = 1, \ldots, P)$ discriminant coefficient can be used to determine whether large or small values of the p^{th} feature variable for an individual in a pair, relative to the values of the same feature variable for the individual's sibling in the same pair, are associated with group 1 compared with group 2, holding all other feature variables fixed. For example, in the context of paired post-mortem tissue studies, we can use the discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ to identify the biomarkers that best distinguish a normal control in a given pair from an individual with schizophrenia in the same pair who is essentially identical with regards to the pairing variables, i.e., age at death, gender, and PMI. Also, the sign of the p^{th} discriminant coefficient can be used to determine whether the p^{th} biomarker is larger or smaller for a normal control compared with an individual with schizophrenia in a given pair, holding all other biomarkers fixed.

An intriguing parallel exists between the adjustment approaches we develop in Sections 3.3.1.1 and 3.3.1.2 when these two approaches are applied to paired data. Specifically, we show in Section 3.3.2.1 that when we apply the linear discriminant rule in (3.14) in the data

setting, we get the same rule as that obtained when we apply the pairwise difference rule in (3.16).

3.3.1.3 Classifying Two Pair Members using Known Pair Effect It is interesting to note that we could have obtained the same classification rule in (3.16) by applying the methodology we develop in Section 3.3.1.1 to the "stacked" feature vector $\mathbf{Y}^+ = \left[\begin{smallmatrix}\mathbf{Y}_{ind}\\ \mathbf{Y}_{sib}\end{smallmatrix}\right]$, where we randomly assign each pair member as an individual or sibling. To elaborate, we can assume that for known $\boldsymbol{\gamma}$, $\mathbf{Y}^+ \sim N_{2P}(\boldsymbol{\mu}_i^+, \boldsymbol{\Sigma}^+)$ in the i^{th} group ordering $(i = 1, 2)$ for a given pair, where $\boldsymbol{\mu}_1^+ = \left[\begin{smallmatrix}\boldsymbol{\mu}_1+\boldsymbol{\gamma}\\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}\end{smallmatrix}\right]$, $\boldsymbol{\mu}_2^+ = \left[\begin{smallmatrix}\boldsymbol{\mu}_2+\boldsymbol{\gamma}\\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}\end{smallmatrix}\right]$, $\boldsymbol{\Sigma}^+ = \left[\begin{smallmatrix}\boldsymbol{\Sigma}+\boldsymbol{\Psi} & \boldsymbol{\Psi}\\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}+\boldsymbol{\Psi}\end{smallmatrix}\right]$, and $\boldsymbol{\Psi}$ represents the covariance between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ such that $\boldsymbol{\Psi}' = \boldsymbol{\Psi}$, a typically reasonable assumption in the context of post-mortem tissue studies. If $\mathbf{Y}^+$ belongs to the 1^{st} group ordering, then $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ come from groups 1 and 2, respectively. Otherwise, $\mathbf{Y}_{ind}$ belongs to group 2 and $\mathbf{Y}_{sib}$ belongs to group 1. The details regarding the computation of the linear discriminant rule based on this model for $\mathbf{Y}^+$ are provided in Appendix A.1.

When we're dealing with paired data, the pairwise difference and stacked approaches we developed in Sections 3.3.1.2 and 3.3.1.3, respectively, can be shown to produce identical linear discriminant rules. This implies that in the paired case, the pairwise difference and stacked approaches succeed in answering our primary question of interest, namely, which of the feature variables of interest best discriminate an individual belonging to group 1 from an individual belonging to group 2 in a given pair. In addition, both approaches can be shown to yield identical classification results.

However, we show in Sections 3.5.1.2 and 3.5.1.3 that in the multiple group case, the stacked and differencing approaches produce not only different classification regions, but also different classification results. More importantly, in the multiple group case, the stacked approach does not appear to answer our primary question of interest, namely, how to identify the set of feature variables that best discriminates among the g groups of interest, once the effect of group matching has been taken into account. In fact, the stacked approach can be shown to produce complex results that are difficult to interpret when we match across more than two groups. Also, in the context of classification trees, the stacked approach fails to produce useful results when we match across any number of groups, as we discuss later on in Section 4.4.1.3.

3.3.2 Normal Populations with Unknown Parameters

We now discuss how to implement the methods we develop in Sections 3.3.1.1 to 3.3.1.3 on training data consisting of $\mathbf{y}_{ik}$, the observed feature vector for the member of the k^{th} pair belonging to group i ($i = 1, 2;\ k = 1, \ldots, K$).

3.3.2.1 Classifying Each Member of a Given Pair, with Unknown Pair Effect

Since we know the feature measurements for an individual and their sibling in each pair in the training data, we can apply our model in Section 3.3.1.1 to the random feature vectors corresponding to each member of each training pair, so that we may estimate $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$, as well as $\boldsymbol{\gamma}$ for each training pair. To clarify, for each of the K pairs in the training data, we let $\mathbf{Y}_{ik}$ denote the random feature vector corresponding to the member of the k^{th} pair belonging to group i ($i = 1, 2;\ k = 1, \ldots, K$), with mean $E[\mathbf{Y}_{ik}] = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Using the training data, we fit the model for $E[\mathbf{Y}_{ik}]$ via ML estimation, which we can show, using standard arguments, yields the family of estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) = \bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot} - \mathbf{c}^*$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*) = \bar{\mathbf{y}}_{\cdot k} + \mathbf{c}^*$, where $\bar{\mathbf{y}}_{i\cdot} = \frac{\sum_{k=1}^{K} \mathbf{y}_{ik}}{K}$, $\bar{\mathbf{y}}_{\cdot k} = \frac{\sum_{i=1}^{2} \mathbf{y}_{ik}}{2}$, $\bar{\mathbf{y}}_{\cdot\cdot} = \frac{\sum_{i=1}^{2}\sum_{k=1}^{K} \mathbf{y}_{ik}}{2K}$, $\mathbf{c}^* = -\bar{\mathbf{y}}_{\cdot\cdot} + \mathbf{c}$, and $\mathbf{c} \in \mathbb{R}^P$. Although the estimate $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)$ is not unique, the ML estimate of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is, which is given by $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*) - \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*) = \bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot}$. In addition, the estimate of $\boldsymbol{\mu}_i + \boldsymbol{\gamma}_k$ is unique and is given by $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) + \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*) = \bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{\cdot\cdot} + \bar{\mathbf{y}}_{\cdot k}$.

The ML estimate of $\boldsymbol{\Sigma}$ is given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{2K}\left[\sum_{i=1}^{2}\sum_{k=1}^{K}(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*))(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*))'\right].$$

By substituting the estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$, we have that $\hat{\boldsymbol{\Sigma}} = \frac{1}{4}\left(2\hat{\boldsymbol{\Sigma}}_D\right) = \frac{1}{2}\hat{\boldsymbol{\Sigma}}_D$, where

$$\hat{\boldsymbol{\Sigma}}_D = \frac{1}{2K}\left[\sum_{k=1}^{K}\left(\mathbf{D}_{1k,y} - \bar{\mathbf{D}}_{1\cdot,y}\right)\left(\mathbf{D}_{1k,y} - \bar{\mathbf{D}}_{1\cdot,y}\right)'\right] = \frac{1}{2K}\left[\sum_{k=1}^{K}\left(\mathbf{D}_{2k,y} - \bar{\mathbf{D}}_{2\cdot,y}\right)\left(\mathbf{D}_{2k,y} - \bar{\mathbf{D}}_{2\cdot,y}\right)'\right],$$

$\mathbf{D}_{ik,y} = \mathbf{y}_{ik} - \mathbf{y}_{jk}$, and $\bar{\mathbf{D}}_{i\cdot,y} = \bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{j\cdot}$ ($i, j = 1, 2;\ i \neq j$). At this point, it is clear that $\hat{\boldsymbol{\Sigma}}$ is unique, as is well known.

Intriguingly, although the parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\gamma}_k$ are not identifiable in our model, the classification regions in (3.14) remain invariant when we apply them to the observations in

the training data, which we can do to estimate the precision of the rule in (3.14). To clarify, we first point out that when applied to the training data, (3.14) is of the form

$$\begin{aligned} R_1: \ & \left[\hat{\tilde{\mathbf{y}}}_{ik} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*) + \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*))\right]' \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*) - \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*)) \geq 0, \\ R_2: \ & \left[\hat{\tilde{\mathbf{y}}}_{ik} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*) + \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*))\right]' \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*) - \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*)) < 0, \end{aligned} \tag{3.17}$$

where $\hat{\tilde{\mathbf{y}}}_{ik} = \mathbf{y}_{ik} - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$ denotes the training feature data that have been adjusted for the effect of pairing. When we substitute the formulas for $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*)$, $\hat{\boldsymbol{\mu}}_2(\mathbf{c}^*)$, $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$, and $\hat{\boldsymbol{\Sigma}}$, (3.17) reduces to

$$R_1: \ \mathbf{D}'_{ik,y}\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot}) \geq 0, \quad R_2: \ \mathbf{D}'_{ik,y}\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot}) < 0. \tag{3.18}$$

The estimated rule in (3.18) could also have been obtained if we had used the unbiased estimate $\hat{\boldsymbol{\Sigma}}_D^* = \frac{2K}{2(K-1)}\hat{\boldsymbol{\Sigma}}_D$. We see that the estimated discriminant coefficient vector $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot})$ is unique, so that the non-identifiability of $\boldsymbol{\mu}_i$ is not an issue if we want to identify the feature variables that best discriminate between an individual from group 1 and an individual from group 2 in a particular pair, which is an important result. For example, in the context of post-mortem tissue studies, the fact that $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot})$ is unique implies that we would identify a singular subset of biomarkers that best distinguishes a normal control from an individual with schizophrenia in any given pair.

Based on the adjusted training feature data $\tilde{\mathbf{y}}_{ik}$, the resubstitution method or K-fold cross validation (where the adjusted training feature measurements $\tilde{\mathbf{y}}_{1k}$ and $\tilde{\mathbf{y}}_{2k}$ for each of the K pairs are omitted at a time) as described in Section 3.2.3 are two methods that can be used to obtain an estimate of the probability of misclassification for the rule in (3.14).

On the other hand, if we want to use the rule in (3.14) to classify an individual in a new pair that is not part of the training data, we must re-estimate $\boldsymbol{\gamma}$ for this pair since $\boldsymbol{\gamma}$ is specific to each pair. In this case, we extend Lachenbruch and Tu et al.'s conditional model to include parameters that must be re-estimated for each new pair, namely, $\boldsymbol{\gamma}$. Specifically, if we know the feature measurements for an individual and their sibling in a particular pair, i.e., $\mathbf{y}_{ind}$ and $\mathbf{y}_{sib}$, we can begin by applying the model in Section 3.3.1.1 to the random feature vectors $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ for both pair members, while using the estimates $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*)$, $\hat{\boldsymbol{\mu}}_2(\mathbf{c}^*)$, and $\hat{\boldsymbol{\Sigma}}$ obtained from the training data. In other words, for a given pair and conditional on

the estimates $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*)$, $\hat{\boldsymbol{\mu}}_2(\mathbf{c}^*)$, and $\hat{\boldsymbol{\Sigma}}$, we let $\mathbf{Y}_{ind} \sim N_P(\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) + \boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$ in the i^{th} group and $\mathbf{Y}_{sib} \sim N_P(\hat{\boldsymbol{\mu}}_j(\mathbf{c}^*)+\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$ in the j^{th} group $(i, j = 1, 2; i \neq j)$, where $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ are assumed to be independent. Based on these distributions, we can alternately consider $\mathbf{Y}_{ind} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) \equiv \mathbf{Y}^*_{ind} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$ and $\mathbf{Y}_{sib} - \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*) \equiv \mathbf{Y}^*_{sib} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$. From the likelihood function based on $\mathbf{y}^*_{ind}$ and $\mathbf{y}^*_{sib}$, it can be shown that the ML estimate of $\boldsymbol{\gamma}$ is given by $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*) = \frac{1}{2}(\mathbf{y}^*_{ind}+\mathbf{y}^*_{sib}) = \frac{1}{2}\left[(\mathbf{y}_{ind} + \mathbf{y}_{sib}) - (\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*) + \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*))\right]$. Since $\mathbf{Y}^*_{ind}$ and $\mathbf{Y}^*_{sib}$ are identically distributed, the corresponding likelihood function remains invariant regardless of which groups $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ belong to, so that our estimate of $\boldsymbol{\gamma}$ is unique. We note that even if $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ are not independent, the likelihood function for $\mathbf{y}^*_{ind}$ and $\mathbf{y}^*_{sib}$ is still invariant, as long as the covariance matrix between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ is symmetric. Once we replace $\boldsymbol{\gamma}$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ in (3.14) with their corresponding estimates, we obtain

$$R_1 : \ (\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.}) \geq 0, \quad R_2 : \ (\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.}) < 0, \tag{3.19}$$

which is the same rule as that provided in (3.16) based on the pairwise difference $\mathbf{y}_{ind} - \mathbf{y}_{sib}$, with the coefficient vector $\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in (3.16) replaced with the estimate $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.})$.

In examining our estimate $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$, we see that if we only knew the observation for an individual in a new pair, we could not have estimated $\boldsymbol{\gamma}$ for this individual in the manner previously described. Without an estimate of the parameter $\boldsymbol{\gamma}$, using the rule in (3.14) to classify this individual is not feasible. Of course, this is intuitively clear in the post-mortem brain tissue setting, since any new brain tissue sample must be processed with the appropriate reagents. The quality of these reagents can be viewed as one major determinant of the value of $\boldsymbol{\gamma}$. For example, suppose that the examined biomarkers tend to be higher in controls compared to schizophrenia subjects and that the processing employed for a new pair collectively elevates the pair's biomarker values above that seen by reagents used for the training data and by an unknown amount. Upon observing that only one member's feature measurements are relatively high, one cannot classify that individual into either the control or schizophrenia diagnostic group without knowing $\boldsymbol{\gamma}$.

3.3.2.2 Classifying Each Member of a Given Pair using Pairwise Feature Difference We can also consider an alternate estimation approach that is based on the model for the pairwise differenced random feature vector $\mathbf{Y}_{ind} - \mathbf{Y}_{sib}$ in Section 3.3.1.2. The parameters $\boldsymbol{\mu}_1^{\text{diff}}$, $\boldsymbol{\mu}_2^{\text{diff}}$, and $\boldsymbol{\Sigma}_*$ in this model can be estimated using ML estimation based on the training feature differences $\mathbf{D}_{ik,y} = \mathbf{y}_{ik} - \mathbf{y}_{jk}$ $(i, j = 1, 2;\ i \neq j;\ k = 1, \ldots, K)$, where $\mathbf{D}_{ik,y}$ is computed for the i^{th} group member of the k^{th} pair in the training data. Referring back to our model for $\mathbf{Y}_{ind} - \mathbf{Y}_{sib}$, we have that the differences $\mathbf{D}_{1k,y}$ belong to the first population, while the differences $\mathbf{D}_{2k,y}$ belong to the second population.

The ML estimates of $\boldsymbol{\mu}_1^{\text{diff}}$, $\boldsymbol{\mu}_2^{\text{diff}}$, and $\boldsymbol{\Sigma}_*$ can be shown to equal, respectively, $\bar{\mathbf{D}}_{1.,y} = \bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.}$, $\bar{\mathbf{D}}_{2.,y} = \bar{\mathbf{y}}_{2.} - \bar{\mathbf{y}}_{1.}$, and $\hat{\boldsymbol{\Sigma}}_D$, the formula of which is given in Section 3.3.2.1. Once we plug these estimates into (3.15), we obtain the same rule as in (3.19). Also, when we apply the rule in (3.15) to the differenced training feature data $\mathbf{D}_{ik,y}$, we get the same classification regions as in (3.18).

Based on the training feature differences $\mathbf{D}_{ik,y}$, resubstitution or K-fold cross validation (where the differences $\mathbf{D}_{1k,y}$ and $\mathbf{D}_{2k,y}$ for each of the K pairs in the training data are omitted at a time) as described in Section 3.1.2 can be used to obtain an estimate of the probability of misclassification for the rule in (3.15).

In conclusion, the rule in (3.14) based on the feature vector $\tilde{\mathbf{y}}_{ind}$ that is adjusted for the effect of pairing is the same as the rule in (3.16) based on the pairwise feature difference $\mathbf{y}_{ind} - \mathbf{y}_{sib}$ when applied in the data setting. Based on this fact, we have that both of the estimation approaches in Sections 3.3.2.1 and 3.3.2.2 not only yield the same classification results, but also help us identify, via the estimated discriminant coefficient vector $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.})$, the same set of feature variables that best discriminates between an individual from group 1 and an individual from group 2 in any given pair.

3.3.2.3 Classifying Both Members of a Given Pair, with Unknown Pair Effect The pairwise difference and stacked approaches we developed in Sections 3.3.1.2 and 3.3.1.3, respectively, were shown to yield the same linear discriminant classification rule in (3.16). Thus, we do not provide the details of how the stacked approach can be implemented using training data because Section 3.3.2.2 handles the training data results. We do point out that using fairly detailed calculations, the stacked approach can be shown to produce the

estimated rule in (3.19) when applied to paired data.

3.4 PAIRED LINEAR DISCRIMINANT ANALYSIS WITH COVARIATES

3.4.1 Normal Populations with Known Parameters

In Section 3.4, we extend the methodology developed for paired LDA in Section 3.3 to account not only for the pairing of individuals, but also for the effects on $\mathbf{Y}$ that are attributed to the covariate vector $\mathbf{X}$. While this section is included for completeness, the details are very similar to Section 3.3, so that the reader can safely move ahead to Section 3.5.

As we did in our development of paired LDA, we begin from a population based perspective and extend in Sections 3.4.1.1 to 3.4.1.3 the procedures we developed in Sections 3.3.1.1 to 3.3.1.3 for paired LDA to also account for the effects of additional covariates. In our discussion, we retain our assumptions from Section 3.3.1 of equal misclassification costs and equal prior probabilities.

It is worth noting that if we choose to ignore the effect of pairing on the feature data, we could apply general covariance adjusted LDA to instead account for the effects of the variables on which individuals were paired. For example, this application corresponds to the secondary models described in Section 2.2 that are used in the analysis of many post-mortem tissue studies conducted under the direction of the CCNMD. When we accounted for pairing, we assumed that the feature vector for each member of a particular pair has prior probability 0.5 of belonging to either group 1 or group 2 because we know that if one member belongs to one group, then his or her sibling must belong to the other group. However, when we ignore pairing, we no longer have this kind of information since we only consider the feature data for one randomly selected individual in the population, rather than the feature data for two individuals in a certain pair. As a result, depending on the context, it may not necessarily be appropriate to assume in the unpaired case that $\mathbf{Y}$ has an equal prior probability of belonging to either of the two groups under consideration. In particular, depending on whether we want to use the model of $\mathbf{Y}$ to classify a randomly chosen individual, we may, for example, want to use the population proportion of normal

controls and that of individuals with schizophrenia.

3.4.1.1 Classifying One Pair Member using Known Pair and Covariate Effects

Let $(\mathbf{Y}_{ind}, \mathbf{X}_{ind})$ denote the random feature and covariate vectors for any individual belonging to a pair. In our extension of the paired adjustment methodology we develop in Section 3.3.1.1, we assume that for known $\boldsymbol{\gamma}$ and given $\mathbf{X}_{ind} = \mathbf{x}_{ind}$, $\mathbf{Y}_{ind} \sim N_P(\boldsymbol{\mu}_i + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{ind}, \boldsymbol{\Sigma}_{(x)})$ in the i^{th} group $(i = 1, 2)$. The conditional variance-covariance matrix $\boldsymbol{\Sigma}_{(x)}$ is assumed to be common to both groups, independent of the value of $\mathbf{x}_{ind}$, and positive definite. In addition, $\boldsymbol{\beta} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,S} \\ \vdots & \vdots & \vdots \\ \beta_{P,1} & \cdots & \beta_{P,S} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_P \end{bmatrix}$ is a known parameter matrix that does not depend on group, where $\beta_{p,s}$ is the parameter for the p^{th} feature variable that corresponds to the s^{th} covariate $(p = 1, \ldots, P; s = 1, \ldots, S)$. Based on these conditional densities for $\mathbf{Y}_{ind}$, we can apply the conditional rule in (3.7) to obtain the following classification regions:

$$\begin{aligned} R_{1(x)} : \left[\tilde{\mathbf{y}}_{ind(x)} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \\ R_{2(x)} : \left[\tilde{\mathbf{y}}_{ind(x)} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right]' \boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0, \end{aligned} \tag{3.20}$$

where $\tilde{\mathbf{y}}_{ind(x)} = \mathbf{y}_{ind} - \boldsymbol{\gamma} - \boldsymbol{\beta}\mathbf{x}_{ind}$ denotes an individual's feature measurement that has been adjusted for both pairing and covariate effects. Using the rule in (3.20), we classify an individual with adjusted feature measurement $\tilde{\mathbf{y}}_{ind(x)}$ into group 1 if this value falls into $R_{1(x)}$ and to group 2 otherwise.

From (3.20), we can use the adjusted linear discriminant function $\tilde{\mathbf{y}}'_{ind(x)}\boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ to identify which of the feature variables of interest best discriminate between groups 1 and 2, once the effects of both pairing and covariates on the feature data have been adjusted for. In addition, we can use the sign of the p^{th} $(p = 1, \ldots, P)$ element of the adjusted discriminant coefficient vector $\boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ to determine whether the p^{th} adjusted feature variable is larger or smaller for group 1 compared with group 2, holding all other adjusted feature variables fixed.

In our discussion, we assume that the conditional mean of the feature vector $\mathbf{Y}_{ind}$ depends on a linear function of the covariate data, namely, $\boldsymbol{\beta}\mathbf{x}_{ind}$. However, we can easily generalize to the case where for a given pair and value of $\mathbf{x}_{ind}$, $\mathbf{Y}_{ind} \sim N_P(\boldsymbol{\mu}_i + \boldsymbol{\gamma} + \boldsymbol{\rho}(\mathbf{x}_{ind}; \boldsymbol{\Theta}), \boldsymbol{\Sigma}_{(x)})$ in the i^{th} group, where the function $\boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$ is defined as in Section 3.2.3.

3.4.1.2 Classifying One Pair Member using Covariate Adjusted Pairwise Feature Difference

An alternate approach we can take is to extend our pairwise differencing approach in Section 3.3.1.2 to also handle covariate effects, which we can do by applying Lachenbruch and Tu et al.'s conditional model to the differenced random feature vector $\mathbf{Y}_{ind} - \mathbf{Y}_{sib}$. To elaborate, we first let $\mathbf{X}_{ind}$ and $\mathbf{X}_{sib}$ denote the random covariate vectors for an individual and their sibling in a pair, and let $\mathbf{X}_{ind} - \mathbf{X}_{sib} \equiv \mathbf{X}_{\text{diff}}$ denote the random differenced covariate vector for an individual in this pair. Given $\mathbf{X}_{\text{diff}} = \mathbf{x}_{\text{diff}}$, we assume $(\mathbf{Y}_{ind} - \mathbf{Y}_{sib}) \sim N_P(\boldsymbol{\mu}_i^{\text{diff}} + \boldsymbol{\beta}\mathbf{x}_{\text{diff}}, \boldsymbol{\Sigma}_{*(x)})$ in the i^{th} population $(i = 1, 2)$, where $\boldsymbol{\mu}_i^{\text{diff}}$ are defined as in Section 3.3.1.2. In addition, $\boldsymbol{\Sigma}_{*(x)}$ remains the same for each of the two populations, regardless of whether the conditional variance-covariance matrix for $\mathbf{Y}_{ind}$ and that of $\mathbf{Y}_{sib}$ are the same in each population or whether the conditional covariance matrix between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ is symmetric.

Retaining our assumption from Section 3.3.1.2 that the prior probability of each population is 0.5, we can apply general covariance adjusted LDA based on the conditional distributions of $\mathbf{Y}_{ind} - \mathbf{Y}_{sib}$ in each population to obtain the following regions:

$$\begin{aligned} R_{1(x)}^{\text{diff}} &: \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib}) - \frac{1}{2}\left(\boldsymbol{\mu}_1^{\text{diff}} + \boldsymbol{\mu}_2^{\text{diff}}\right)\right]' \boldsymbol{\Sigma}_{*(x)}^{-1}\left(\boldsymbol{\mu}_1^{\text{diff}} - \boldsymbol{\mu}_2^{\text{diff}}\right) \geq 0, \\ R_{2(x)}^{\text{diff}} &: \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib}) - \frac{1}{2}\left(\boldsymbol{\mu}_1^{\text{diff}} + \boldsymbol{\mu}_2^{\text{diff}}\right)\right]' \boldsymbol{\Sigma}_{*(x)}^{-1}\left(\boldsymbol{\mu}_1^{\text{diff}} - \boldsymbol{\mu}_2^{\text{diff}}\right) < 0. \end{aligned} \tag{3.21}$$

After further simplification, the regions in (3.21) can be re-expressed as

$$\begin{aligned} R_{1(x)}^{\text{diff}} &: \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})\right]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \\ R_{2(x)}^{\text{diff}} &: \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})\right]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0. \end{aligned} \tag{3.22}$$

For each individual in a given pair, we compute the covariate adjusted pairwise feature difference $(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})$. If this difference falls into region $R_{1(x)}^{\text{diff}}$, then it is assigned to the first population and we classify that individual into the first group. Otherwise, if $(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})$ falls into region $R_{2(x)}^{\text{diff}}$, then this difference is assigned to the second population and we classify that individual into the second group. As was the case for the pairwise difference rule in (3.16), we have that if an individual in a pair is classified into the first group based on the rule in (3.22), then their sibling is classified into the second group, and vice versa.

We note that the covariate adjusted pairwise difference $(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})$ can also be written as $(\mathbf{y}_{ind} - \boldsymbol{\beta}\mathbf{x}_{ind}) - (\mathbf{y}_{sib} - \boldsymbol{\beta}\mathbf{x}_{sib})$, which is the difference between the covariate adjusted feature measurement for an individual in a pair and that of their sibling. When we view the covariate adjusted difference in this manner, we then have that the adjusted discriminant coefficient vector $\boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ can be used to identify which of the feature variables of interest, once they've been suitably adjusted for covariate effects, best distinguish an individual belonging to group 1 from that belonging to group 2 in a given pair. For example, in the context of post-mortem tissue studies, this coefficient vector can be used to identify the subset of biomarkers, once the effects of covariates such as brain tissue storage time are adjusted for, that best discriminates a normal control in a pair from an individual with schizophrenia in the same pair. The sign of the p^{th} $(p = 1, \ldots, P)$ coefficient can then be used to determine whether large or small values of the p^{th} covariate adjusted feature variable for an individual in a pair, relative to the values of the same covariate adjusted feature variable for the individual's sibling in the same pair, are associated with group 1 compared with group 2, holding all other covariate adjusted feature variables fixed.

3.4.1.3 Classifying Two Pair Members using Known Pair and Covariate Effects

The linear discriminant rule in (3.22) could also have been obtained if we had applied our methodology in Section 3.4.1.1 to the stacked feature vector $\mathbf{Y}^+ = \left[\begin{smallmatrix}\mathbf{Y}_{ind}\\ \mathbf{Y}_{sib}\end{smallmatrix}\right]$. To clarify, we first define $\mathbf{X}^+ = \left[\begin{smallmatrix}\mathbf{X}_{ind}\\ \mathbf{X}_{sib}\end{smallmatrix}\right]$ as the random covariate vector corresponding to an individual and their sibling in a pair. Given $\mathbf{X}^+ = \mathbf{x}^+$, we assume $\mathbf{Y}^+ \sim N_{2P}(\boldsymbol{\mu}_{i(x)}^+, \boldsymbol{\Sigma}_{(x)}^+)$ in the i^{th} group ordering $(i = 1, 2)$ for a given pair, where $\boldsymbol{\mu}_{1(x)}^+ = \left[\begin{smallmatrix}\boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib}\end{smallmatrix}\right]$, $\boldsymbol{\mu}_{2(x)}^+ = \left[\begin{smallmatrix}\boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib}\end{smallmatrix}\right]$, $\boldsymbol{\Sigma}_{(x)}^+ = \left[\begin{smallmatrix}\boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi} & \boldsymbol{\Psi}\\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi}\end{smallmatrix}\right]$, and $\boldsymbol{\Psi}$ is defined as in Section 3.3.1.3. The details regarding the computation of the linear discriminant classification rule based on the conditional model for $\mathbf{Y}^+$ are provided in Appendix A.2.

3.4.2 Normal Populations with Unknown Parameters

In the next three sections, we discuss how to implement the adjustment procedures we develop in Sections 3.4.1.1 to 3.4.1.3 based on training data consisting of $(\mathbf{y}_{ik}, \mathbf{x}_{ik})$, where $\mathbf{y}_{ik}$ is defined as in Section 3.3.2 and $\mathbf{x}_{ik}$ denotes the observed covariate vector for the member of the k^{th} pair belonging to group i $(i = 1, 2;\ k = 1, \ldots, K)$.

3.4.2.1 Classifying Each Member of a Given Pair, with Unknown Pair and Covariate Effects

We first discuss how our conditional model in Section 3.4.1.1 can be used to estimate, using the available training data, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}_i$ $(i = 1, 2)$, and $\boldsymbol{\Sigma}_{(x)}$ needed to compute the classification rule in (3.20).

To clarify, we define $\mathbf{Y}_{ik}$ as in Section 3.3.2.1, with conditional mean $E[\mathbf{Y}_{ik}|\mathbf{x}_{ik}] = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_k + \boldsymbol{\beta}\mathbf{x}_{ik}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{(x)}$. Based on the training data, we use ML estimation to fit the model for $E[\mathbf{Y}_{ik}|\mathbf{x}_{ik}]$, whose design matrix we assume satisfies certain conditions that ensure that the ML estimate $\hat{\boldsymbol{\beta}}$ is unique. In fitting this model, we obtain the family of estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) = \bar{\mathbf{y}}_{i.} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{i.} - (\bar{\mathbf{y}}_{..} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{..}) - \mathbf{c}^*_{\mathbf{x}}$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) = \bar{\mathbf{y}}_{.k} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{.k} + \mathbf{c}^*_{\mathbf{x}}$, where $\bar{\mathbf{y}}_{i.}$, $\bar{\mathbf{y}}_{.k}$, and $\bar{\mathbf{y}}_{..}$ are defined as in Section 3.3.2.1, $\bar{\mathbf{x}}_{i.} = \frac{\sum_{k=1}^{K}\mathbf{x}_{ik}}{K}$, $\bar{\mathbf{x}}_{.k} = \frac{\sum_{i=1}^{2}\mathbf{x}_{ik}}{2}$, $\bar{\mathbf{x}}_{..} = \frac{\sum_{i=1}^{2}\sum_{k=1}^{K}\mathbf{x}_{ik}}{2K}$, and $\mathbf{c}^*_{\mathbf{x}} = -(\bar{\mathbf{y}}_{..} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{..}) + \mathbf{c}$. The ML estimate of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is then given by $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}})$ $= \bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.} - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{1.} - \bar{\mathbf{x}}_{2.})$, which we see is unique.

The ML estimate of $\boldsymbol{\Sigma}_{(x)}$ is equal to

$$\hat{\boldsymbol{\Sigma}}_{(x)} = \frac{1}{2K}\left[\sum_{i=1}^{2}\sum_{k=1}^{K}(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik})(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik})'\right].$$

When we substitute the estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}})$, we have that $\hat{\boldsymbol{\Sigma}}_{(x)} = \frac{1}{4}\left(2\hat{\boldsymbol{\Sigma}}_{D(x)}\right) = \frac{1}{2}\hat{\boldsymbol{\Sigma}}_{D(x)}$, where

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{D(x)} &= \frac{1}{2K}\left[\sum_{k=1}^{K}\left(\hat{\mathbf{D}}^{adj}_{1k,y} - \bar{\hat{\mathbf{D}}}^{adj}_{1.,y}\right)\left(\hat{\mathbf{D}}^{adj}_{1k,y} - \bar{\hat{\mathbf{D}}}^{adj}_{1.,y}\right)'\right] \\ &= \frac{1}{2K}\left[\sum_{k=1}^{K}\left(\hat{\mathbf{D}}^{adj}_{2k,y} - \bar{\hat{\mathbf{D}}}^{adj}_{2.,y}\right)\left(\hat{\mathbf{D}}^{adj}_{2k,y} - \bar{\hat{\mathbf{D}}}^{adj}_{2.,y}\right)'\right],\end{aligned}$$

$\hat{\mathbf{D}}^{adj}_{ik,y} = \mathbf{D}_{ik,y} - \hat{\boldsymbol{\beta}}\mathbf{D}_{ik,x}$, $\mathbf{D}_{ik,x} = \mathbf{x}_{ik} - \mathbf{x}_{jk}$ $(i, j = 1, 2;\ i \neq j)$, and $\bar{\hat{\mathbf{D}}}^{adj}_{i.,y} = \bar{\mathbf{D}}_{i.,y} - \hat{\boldsymbol{\beta}}\bar{\mathbf{D}}_{i.,x} =$ $(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{j.}) - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{j.})$.

Despite the fact that $\boldsymbol{\mu}_i$ and $\boldsymbol{\gamma}_k$ are not identifiable, the classification regions in (3.20) remain the same when applied to the training data, regardless of the value of $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}})$. In applying this rule, (3.20) takes on the form

$$\begin{aligned}R_{1(x)}:&\ \left[\hat{\bar{\mathbf{y}}}_{ik(x)} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}}) + \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}}))\right]'\hat{\boldsymbol{\Sigma}}^{-1}_{(x)}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}})) \geq 0, \\ R_{2(x)}:&\ \left[\hat{\bar{\mathbf{y}}}_{ik(x)} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}}) + \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}}))\right]'\hat{\boldsymbol{\Sigma}}^{-1}_{(x)}(\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}})) < 0,\end{aligned} \tag{3.23}$$

where $\hat{\tilde{\mathbf{y}}}_{ik(x)} = \mathbf{y}_{ik} - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik}$ denotes the training feature data that have been adjusted for pairing and covariate effects. Based on the formulas for $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}})$, $\hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}})$, $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}})$, and $\hat{\boldsymbol{\Sigma}}_{(x)}$, the rule in (3.23) can be expressed as

$$\begin{aligned} R_{1(x)}: \ & \left[\mathbf{D}_{ik,y} - \hat{\boldsymbol{\beta}}\mathbf{D}_{ik,x}\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{D(x)} \left[\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot} - \hat{\boldsymbol{\beta}}\left(\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot}\right)\right] \geq 0, \\ R_{2(x)}: \ & \left[\mathbf{D}_{ik,y} - \hat{\boldsymbol{\beta}}\mathbf{D}_{ik,x}\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{D(x)} \left[\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot} - \hat{\boldsymbol{\beta}}\left(\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot}\right)\right] < 0, \end{aligned} \tag{3.24}$$

which could also have obtained by using the unbiased estimate $\hat{\boldsymbol{\Sigma}}^*_{D(x)} = \frac{2K}{2(K-1)}\hat{\boldsymbol{\Sigma}}_{D(x)}$. To estimate the probability of misclassification associated with the conditional rule in (3.20), we can use either resubstitution or K-fold cross validation as described in Section 3.3.2.1 based on the adjusted training feature measurements $\tilde{\mathbf{y}}_{ik(x)}$.

In order to use the rule in (3.20) to classify an individual in a pair that is not part of the training data, the parameter $\boldsymbol{\gamma}$ must be re-estimated for this pair, as was the case when we only accounted for the effect of pairing. If we know the feature and covariate measurements for both pair members, i.e., $(\mathbf{y}_{ind}, \mathbf{x}_{ind})$ and $(\mathbf{y}_{sib}, \mathbf{x}_{sib})$, we can start off by applying our conditional model in Section 3.4.1.1 to the two random feature vectors $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ corresponding to this pair, while using the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$ $(i = 1, 2)$, and $\hat{\boldsymbol{\Sigma}}_{(x)}$ from the training data. In other words, conditional on $\mathbf{x}_{ind}$, $\mathbf{x}_{sib}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}})$, $\hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}})$, and $\hat{\boldsymbol{\Sigma}}_{(x)}$, we assume $\mathbf{Y}_{ind} \sim N_P(\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{ind}, \hat{\boldsymbol{\Sigma}}_{(x)})$ in the i^{th} group and $\mathbf{Y}_{sib} \sim N_P(\hat{\boldsymbol{\mu}}_j(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{sib}, \hat{\boldsymbol{\Sigma}}_{(x)})$ in the j^{th} group for a given pair $(i, j = 1, 2; i \neq j)$, where we assume $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ are independent. From these distributions, we can also consider $\mathbf{Y}_{ind} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ind} \equiv \mathbf{Y}^{*(x)}_{ind} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}}_{(x)})$ and $\mathbf{Y}_{sib} - \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{sib} \equiv \mathbf{Y}^{*(x)}_{sib} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}}_{(x)})$. Based on the resulting likelihood function, the ML estimate of $\boldsymbol{\gamma}$ is equal to $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}}) = \frac{1}{2}(\mathbf{y}^{*(x)}_{ind} + \mathbf{y}^{*(x)}_{sib}) = \frac{1}{2}\left[(\mathbf{y}_{ind} + \mathbf{y}_{sib}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} + \mathbf{x}_{sib}) - (\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*_{\mathbf{x}}) + \hat{\boldsymbol{\mu}}_2(\mathbf{c}^*_{\mathbf{x}}))\right]$. The fact that $\mathbf{Y}^{*(x)}_{ind}$ and $\mathbf{Y}^{*(x)}_{sib}$ are identically distributed implies that the likelihood function based on $\mathbf{y}^{*(x)}_{ind}$ and $\mathbf{y}^{*(x)}_{sib}$ remains the same regardless of the groups to which $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ belong and, thus, the estimate $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}})$ is unique. As long as the covariance matrix between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ is symmetric, this likelihood function remains invariant even if $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ are not independent.

After we plug in the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$ $(i = 1, 2)$, $\hat{\boldsymbol{\Sigma}}_{(x)}$, and $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}})$, the rule in (3.20) becomes

$$\begin{aligned} R_{1(x)}: \ & \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} - \mathbf{x}_{sib})\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{D(x)} \left[\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot} - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})\right] \geq 0, \\ R_{2(x)}: \ & \left[(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} - \mathbf{x}_{sib})\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{D(x)} \left[\bar{\mathbf{y}}_{1\cdot} - \bar{\mathbf{y}}_{2\cdot} - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{1\cdot} - \bar{\mathbf{x}}_{2\cdot})\right] < 0, \end{aligned} \tag{3.25}$$

which we see is the classification rule in (3.22), with $\boldsymbol{\beta}$ and the coefficient vector $\boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ replaced with the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}_{D(x)}^{-1}\left[\bar{\mathbf{y}}_{1.}-\bar{\mathbf{y}}_{2.}-\hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{1.}-\bar{\mathbf{x}}_{2.})\right]$, respectively.

3.4.2.2 Classifying Each Member of a Given Pair using Covariate Adjusted Pairwise Difference Alternately, we can implement our differencing approach in Section 3.4.1.2 using the available training data. First, we let $\mathbf{D}_{ik,Y} \equiv \mathbf{Y}_{ik}-\mathbf{Y}_{jk}$ $(i,j=1,2;\ i\neq j)$ denote the random differenced feature vector for the member of the k^{th} pair belonging to group i $(i=1,2;\ k=1,\ldots,K)$, i.e., in the k^{th} pair, $\mathbf{D}_{ik,Y}$ corresponds to the i^{th} population. Based on the conditional model for the feature difference $\mathbf{Y}_{ind}-\mathbf{Y}_{sib}$ that we specify in Section 3.4.1.2, we assume $\mathbf{D}_{ik,Y}$ has conditional mean $E[\mathbf{D}_{ik,Y}|\mathbf{D}_{ik,x}]=\boldsymbol{\mu}_i^{\text{diff}}+\boldsymbol{\beta}\mathbf{D}_{ik,x}$ and conditional variance-covariance matrix $\boldsymbol{\Sigma}_{*(x)}$. Using the differences $(\mathbf{D}_{ik,y},\mathbf{D}_{ik,x})$, which are defined as in Sections 3.3.2.1 and 3.4.2.1, we use ML estimation to fit the model for $E[\mathbf{D}_{ik,Y}|\mathbf{D}_{ik,x}]$, whose design matrix we assume satisfies certain conditions that ensure that the ML estimate $\hat{\boldsymbol{\beta}}$ is unique. In doing so, we obtain the estimates $\hat{\boldsymbol{\mu}}_i^{\text{diff}}=\bar{\hat{\mathbf{D}}}_{i.,y}^{adj}$ $(i=1,2)$ and $\hat{\boldsymbol{\Sigma}}_{*(x)}=\hat{\boldsymbol{\Sigma}}_{D(x)}$, where $\bar{\hat{\mathbf{D}}}_{i.,y}^{adj}$ and $\hat{\boldsymbol{\Sigma}}_{D(x)}$ are defined as in Section 3.4.2.1. In plugging the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_1^{\text{diff}}$, $\hat{\boldsymbol{\mu}}_2^{\text{diff}}$, and $\hat{\boldsymbol{\Sigma}}_{*(x)}$ into (3.21), we obtain the same rule as in (3.25). We also note that when we apply the rule in (3.21) to the training data, we get the same classification regions as in (3.24).

Using the covariate adjusted training feature differences $\mathbf{D}_{ik,y}-\hat{\boldsymbol{\beta}}\mathbf{D}_{ik,x}$, resubstitution or K-fold cross validation (where the covariate adjusted differences $\mathbf{D}_{1k,y}-\hat{\boldsymbol{\beta}}\mathbf{D}_{1k,x}$ and $\mathbf{D}_{2k,y}-\hat{\boldsymbol{\beta}}\mathbf{D}_{2k,x}$ for each of the K pairs in the training data are omitted one at a time) as described in Section 3.2.3 can be used to estimate the probability of misclassification associated with the conditional rules in (3.21) and, equivalently, (3.22).

Based on the results we obtain in Sections 3.4.2.1 and 3.4.2.2, we have that both of the estimation approaches in these two sections produce the same estimated classification rules and, thus, the same classification results when applied to paired data. They also help us identify, using the estimated adjusted discriminant coefficient vector $\hat{\boldsymbol{\Sigma}}_{D(x)}^{-1}\left[\bar{\mathbf{y}}_{1.}-\bar{\mathbf{y}}_{2.}-\hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{1.}-\bar{\mathbf{x}}_{2.})\right]$, the same set of feature variables that best distinguishes an individual in group 1 from an individual in group 2 in any given pair, once these feature variables have been suitably adjusted for covariate effects.

3.4.2.3 Classifying Both Members of a Given Pair, with Unknown Pair and Covariate Effects

As was the case when we only adjusted for the effect of pairing, it can be shown that when applied in the data setting, our stacked approach in Section 3.4.1.3 yields the same estimated linear discriminant rule in (3.25) based on the covariate adjusted pairwise feature differences.

3.5 ACCOUNTING FOR EFFECT OF MULTIPLE GROUP MATCHING IN LDA

3.5.1 Normal Populations with Known Parameters

We now extend the three adjustment methodologies we develop for pairing in Sections 3.3.1.1 to 3.3.1.3 to handle the case where individuals are matched across $g > 2$ groups, where we again begin our discussion from a population based perspective. Recall that in the paired case, the three methodological approaches yielded linear discriminant rules that were relatively clear and easy to interpret, regardless of whether we wanted to find the most discriminatory subset of feature variables or classify each member belonging to a new pair. On the other hand, the interpretation of the linear discriminant rules we obtain from extending the paired approaches in Sections 3.3.1.1 to 3.3.1.3 to handle matching across more than two groups requires considerably more care, as we show in Sections 3.5.1.1 to 3.5.1.3. To better clarify the results of our three extended adjustment approaches, we briefly describe how these three approaches could be applied to the post-mortem brain biomarker data (Konopaske et al.) discussed in Section 2.4, which dealt with subject matching across three treatment groups. In Section 5.2, we give a detailed discussion of the results we obtain when we actually implement these adjustment approaches in Sections 3.5.1.1 to 3.5.1.3 using the Konopaske data.

Given the feature measurements for all g members of a match, we know that the first member belongs to group i_1, the second member belongs to group i_2, ..., and the g^{th} member belongs to group i_g $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g)$. In this case, it is equally likely that each member belongs to one of the g groups under consideration since we

assume there is no preference for which member is designated as the first member, the second member, ..., or the g^{th} member. Therefore, even though our adjustment methodology can handle any π_i in general, it is appropriate to assume for each match that the feature vector for each member has probability $1/g$ of belonging to any of the g groups, i.e., $\pi_i = 1/g$ $(i = 1, \ldots, g)$. We also retain our assumption of equal misclassification costs.

3.5.1.1 Classifying One Member of a Match using Known Match Effect For known $\boldsymbol{\gamma}$, we assume $\mathbf{Y}_{ind} \sim N_P(\boldsymbol{\mu}_i + \boldsymbol{\gamma}, \boldsymbol{\Sigma})$ in the i^{th} group $(i = 1, \ldots, g)$, where the random feature vector $\mathbf{Y}_{ind}$ corresponds to any individual that belongs to a particular match. In the spirit of Lachenbruch and Tu et al., we apply LDA for multiple groups, as first introduced by Fisher [9] and as described in greater detail, for example, by Anderson [1] and by McLachlan [23], based on the densities of $\mathbf{Y}_{ind}$ to obtain the rule

$$R_i : \left[\tilde{\mathbf{y}}_{ind} - \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)\right]' \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) > 0, \quad j = 1, \ldots, g;\ j \neq i, \tag{3.26}$$

where $\tilde{\mathbf{y}}_{ind} = \mathbf{y}_{ind} - \boldsymbol{\gamma}$ is the feature vector that has been adjusted for the effect of group matching. Using (3.26), we classify an individual in a match with adjusted feature measurement $\tilde{\mathbf{y}}_{ind}$ into the i^{th} group if $\tilde{\mathbf{y}}_{ind}$ falls into region R_i $(i = 1, \ldots, g)$.

Intuitively, (3.26) says that we classify an observation into group i if all $g - 1$ pairwise comparisons of group i versus group j $(i, j = 1, \ldots, g;\ j \neq i)$ indicate that the observation should be classified into group i. Interpreting (3.26) we see that the discriminant function which differentiates group i from group j, while accounting for the effect of group matching, is given by $\tilde{\mathbf{y}}'_{ind}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$. Specifically, once its elements have been suitably standardized, the discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ can be used to identify the feature variables that best discriminate between groups i and j, after adjusting for the effect of group matching. The sign of the p^{th} element $(p = 1, \ldots, P)$ of the coefficient vector $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ can then be used to determine whether the p^{th} adjusted feature variable is larger or smaller for group i compared with group j, holding all other adjusted feature variables fixed.

For example, for the Konopaske biomarker data, suppose we label the haloperidol, olanzapine, and sham treatments as groups 1, 2, and 3, respectively, noting that this group assignment is completely arbitrary. In considering this biomarker data, the three discriminant coefficient vectors $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)$, and $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)$ can help us determine

which of the biomarkers under consideration, once they've been adjusted for the effect of matching, best discriminate between the haloperidol and olanzapine treatment groups, which of them best discriminate between the haloperidol and sham treatment groups, and which of them best discriminate between the olanzapine and sham treatment groups, respectively. In addition, the sign of the p^{th} coefficient of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)$, and $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)$ can be used to determine whether the p^{th} adjusted biomarker is larger or smaller for the haloperidol group compared with the olanzapine group, the haloperidol group compared with the sham group, and the olanzapine group compared with the sham group, respectively, holding all other adjusted biomarker values fixed. Thus, to classify an individual into the sham group, for example, the two comparisons of haloperidol versus sham and olanzapine versus sham should both classify this individual as belonging to the sham group.

3.5.1.2 Classifying One Member of a Match using Feature Difference An alternate approach we can take to account for the effect of group matching on the feature data is to implement traditional LDA for multiple groups on the differenced random feature vector $\mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}$, where $\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$ denote the random feature vectors for an individual and their $g-1$ siblings in a given match. Under this approach, we have that $(\mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}) \equiv \mathbf{Y}_{\text{diff}} \sim N_P(\boldsymbol{\mu}_i^{\text{diff}}, \boldsymbol{\Sigma}_*)$ in the i^{th} population $(i = 1, \ldots, g)$, where $\boldsymbol{\mu}_i^{\text{diff}} = \boldsymbol{\mu}_i - \sum_{\substack{l=1 \\ l\neq i}}^{g}\boldsymbol{\mu}_l$ $(i = 1, \ldots, g)$. In the i^{th} population, $\mathbf{Y}_{ind}$ belongs to the i^{th} group. Recall that in the paired case, the variance-covariance matrix of the difference $\mathbf{Y}_{ind} - \mathbf{Y}_{sib}$ remained the same in each population, even if the variance-covariance matrices of $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ did not stay the same across the two populations. However, when we're dealing with matching across more than two groups, $\boldsymbol{\Sigma}_*$ remains the same in each population if we have that the g variance-covariance matrices for $\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$ and the covariance between any pair of the g feature vectors $\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$ remain the same in each population. We note that the covariance matrix between any pair of these g feature vectors need not be symmetric in order for $\boldsymbol{\Sigma}_*$ to stay the same in each population.

As we stated in Section 3.5.1, each member of a match is equally likely to belong to one of the g groups and the labeling of a member as an individual or as any of the $g-1$ siblings is completely random. Based on these two facts, and the fact that the feature vector $\mathbf{Y}_{ind}$ for an individual in a given match belongs to the i^{th} group in the i^{th} population of $\mathbf{Y}_{\text{diff}}$, we

assume that the prior probability of each population is $1/g$. In the case of multiple groups, applying the rule used for traditional LDA to the difference $\mathbf{Y}_{\text{diff}}$ yields the classification regions

$$R_i^{\text{diff}}: \left[\mathbf{y}_{\text{diff}} - \frac{1}{2}\left(\boldsymbol{\mu}_i^{\text{diff}} + \boldsymbol{\mu}_j^{\text{diff}}\right)\right]' \boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\mu}_i^{\text{diff}} - \boldsymbol{\mu}_j^{\text{diff}}\right) > 0, \quad j = 1, \ldots, g;\ j \neq i. \tag{3.27}$$

For each member of a match, we compute the difference $\mathbf{y}_{\text{diff}} = \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{y}_{sib,m}$, the difference between the feature measurement for that member and the *average* of the feature measurements for their siblings in that match. If this difference falls into region R_i^{diff}, we classify that member into the i^{th} group ($i = 1, \ldots, g$). With the formula for $\boldsymbol{\mu}_i^{\text{diff}}$, the rule in (3.27) can be further simplified as

$$R_i^{\text{diff}}: \left[\mathbf{y}_{\text{diff}} - \frac{g-2}{g-1}\left(\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{1}{g-2}\sum_{\substack{l=1 \\ l\neq i,j}}^{g}\boldsymbol{\mu}_l\right)\right]' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) > 0, \quad j = 1, \ldots, g;\ j \neq i. \tag{3.28}$$

From the rule in (3.28), we can use the discriminant function $\mathbf{y}'_{\text{diff}}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ to identify the feature variables that best discriminate between groups i and j, once the effect of matching on these feature variables has been taken into account. For example, with regards to the Konopaske data, we can compute the difference $\mathbf{y}_{\text{diff}}$ for each monkey in a particular triad and use the rule in (3.28) to classify this monkey based on their value of $\mathbf{y}_{\text{diff}}$. In addition, suppose we retain the same group assignments for the haloperidol, olanzapine, and sham treatment groups as in Section 3.5.1.1. The three discriminant coefficient vectors $\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)$, and $\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)$ allow us to identify the biomarkers that best discriminate between the haloperidol and olanzapine treatment groups, between the haloperidol and sham treatment groups, and between the olanzapine and sham treatment groups, once we account for the effect of triad matching on these biomarkers.

Similar to what we saw in the paired case, we show in Section 3.5.2.1 that when the linear discriminant rule in (3.26) is applied in the data setting, we get the same rule as that obtained when we apply the feature difference rule in (3.28).

3.5.1.3 Classifying All Members of a Match using Known Match Effect The adjustment methodology we develop in Section 3.5.1.1 can also be applied to the stacked feature vector $\mathbf{Y}^+$, which corresponds to all members of a particular match. While the derivation for any number of groups is computationally feasible, we present only the case of matching across three groups for notational convenience.

Let $\mathbf{Y}^+ = \begin{bmatrix} \mathbf{Y}_{ind} \\ \mathbf{Y}_{sib,1} \\ \mathbf{Y}_{sib,2} \end{bmatrix}$ denote the random feature vector corresponding to an individual and their two siblings in a triad and $\mathbf{y}^+$ its observed counterpart, where the assignment of each triad member as an individual or as any one of the two siblings is completely random with no other preferences. Conceptually, based upon the three groups to which each member can belong, $\mathbf{Y}^+$ can belong to one of $3! = 6$ possible group orderings. For known $\boldsymbol{\gamma}$, we assume $\mathbf{Y}^+ \sim N_{3P}(\boldsymbol{\mu}_i^+, \boldsymbol{\Sigma}^+)$ in the i^{th} group ordering $(i = 1, \ldots, 6)$ for a given match, where

$$\boldsymbol{\mu}_1^+ = \begin{bmatrix} \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_2^+ = \begin{bmatrix} \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_3^+ = \begin{bmatrix} \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_4^+ = \begin{bmatrix} \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \end{bmatrix},$$
$$\boldsymbol{\mu}_5^+ = \begin{bmatrix} \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_6^+ = \begin{bmatrix} \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\Sigma}^+ = \begin{bmatrix} \boldsymbol{\Sigma}+\boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}+\boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Sigma}+\boldsymbol{\Psi} \end{bmatrix},$$

and $\boldsymbol{\Psi}$ represents the covariance between any two of the g random feature vectors in that match such that $\boldsymbol{\Psi}' = \boldsymbol{\Psi}$. From this model for $\mathbf{Y}^+$, we can obtain a set of classification regions that allows us to simultaneously assign $\mathbf{y}_{ind}$, $\mathbf{y}_{sib,1}$, and $\mathbf{y}_{sib,2}$ to one of the six possible group orderings. For example, if $\mathbf{y}^+$ were classified into the first group ordering, then $\mathbf{y}_{ind}$, $\mathbf{y}_{sib,1}$, and $\mathbf{y}_{sib,2}$ would be classified into groups 1, 2, and 3, respectively. The details for constructing these classification regions are provided in Appendix B.1.

In the paired case, we showed that the linear discriminant rules obtained from the pairwise difference approach and stacked approach in Sections 3.3.1.2 and 3.3.1.3, respectively, were identical, which implied that both approaches provided the same information regarding the discriminatory ability of the feature data and the same classification results. On the other hand, when we extend to the case of matching across multiple groups, we see a substantial difference in the results obtained when we construct our linear discriminant rule based on the difference $\mathbf{y}_{\text{diff}} \equiv \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{y}_{sib,m}$ and the stacked feature vector $\mathbf{y}^+$. To better illustrate this fact, we briefly review the application of our differenced and stacked approaches to the Konopaske biomarker data.

Recall that if we apply the rule in (3.28) to the biomarker difference $\mathbf{y}_{\text{diff}}$ for each monkey in a specific triad, then we would examine a total of three classification regions, each of which consists of two discriminant functions, that we use to classify this monkey into one of the three drug treatment groups. However, as we show in Appendix B.1, when we consider the stacked biomarker vector $\mathbf{y}^+$ for a matched monkey triad, we examine a total of six classification regions, each consisting of five discriminant functions d_{ij} that are used to simultaneously classify all monkeys in this triad into one of the six possible treatment group orderings, and where some of the 30 possible discriminant functions d_{ij} are distinct and some are the same. For example, retaining the same group assignments for the three treatment groups as in Section 3.5.1.1, we have that classification of a monkey triad into the first group ordering entails classifying the first, second, and third monkeys in that triad into the haloperidol, olanzapine, and sham treatment groups, respectively. Also, classification of a triad into the fourth group ordering means classifying the first, second, and third monkeys in the triad into the olanzapine, sham, and haloperidol treatment groups, respectively. In fact, not only do the linear discriminant rules based on $\mathbf{y}_{\text{diff}}$ and $\mathbf{y}^+$ differ, but they also do not provide us with the same classification information and can be shown to yield different classification results.

Also, from (3.28), we could use the linear discriminant functions $\mathbf{y}'_{\text{diff}}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\mathbf{y}'_{\text{diff}}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3)$, and $\mathbf{y}'_{\text{diff}}\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3)$ to identify the biomarkers that best discriminate between the haloperidol and olanzapine groups, between the haloperidol and sham groups, and between the olanzapine and sham groups, respectively, once we've accounted for the effect of matching on the biomarkers of interest. On the other hand, when we examine all 30 discriminant functions d_{ij} across the six classification regions in our consideration of $\mathbf{y}^+$, it is considerably more difficult to interpret the information that these functions convey from a discrimination viewpoint. To elaborate, we first note that based on our discussion in Appendix B.1, it can be shown that whether or not the discriminant function d_{ij} is positive is what distinguishes the i^{th} group ordering from the j^{th} group ordering $(i, j = 1, \ldots, 6;\ i \neq j)$. For example, whether or not d_{14} is positive is what differentiates between the first and fourth treatment group orderings. In other words, if d_{14} is positive, then the first monkey in a triad is assigned to the haloperidol group as opposed to the olanzapine group, the second monkey is assigned to the olanzapine group as opposed to the sham group, and the third monkey is

assigned to the haloperidol group as opposed to the sham group. However, it is not evident that we can use this information to identify among the biomarkers under consideration those that best discriminate among the three treatment groups once the effect of group matching has been accounted for, which is our main interest in our analysis.

3.5.2 Normal Populations with Unknown Parameters

We now discuss how to implement our adjustment procedures in Sections 3.5.1.1 to 3.5.1.3 using training data consisting of $\mathbf{y}_{ik}$, the observed feature vector for the member of the k^{th} match belonging to the i^{th} group ($i = 1, \ldots, g;\ k = 1, \ldots, K$).

3.5.2.1 Classifying Each Member of a Given Match, with Unknown Match Effect Based on our knowledge of the feature measurements for an individual and their $g-1$ siblings in each match in the training data, we now apply our model in Section 3.5.1.1 to the random feature vectors corresponding to each member of each training match so that we may estimate all parameters in (3.26), including $\boldsymbol{\gamma}$ for each training match.

In other words, we first let $\mathbf{Y}_{ik}$ denote the random feature vector corresponding to the member of the k^{th} match belonging to the i^{th} group ($i = 1, \ldots, g;\ k = 1, \ldots, K$). Based on our assumed model in Section 3.5.1.1, we have that $\mathbf{Y}_{ik}$ has mean $E[\mathbf{Y}_{ik}] = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Once we fit this model using ML estimation, we obtain the family of estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) = \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..} - \mathbf{c}^*$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*) = \bar{\mathbf{y}}_{.k} + \mathbf{c}^*$, where $\bar{\mathbf{y}}_{i.} = \frac{\sum_{k=1}^{K} \mathbf{y}_{ik}}{K}$, $\bar{\mathbf{y}}_{.k} = \frac{\sum_{i=1}^{g} \mathbf{y}_{ik}}{g}$, $\bar{\mathbf{y}}_{..} = \frac{\sum_{i=1}^{g}\sum_{k=1}^{K} \mathbf{y}_{ik}}{gK}$ and $\mathbf{c}^* = -\bar{\mathbf{y}}_{..} + \mathbf{c}$. Despite the fact that $\boldsymbol{\mu}_i$ is not identifiable in our model, the ML estimate of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ is unique and is given by $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*) = \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{j.}$ ($i, j = 1, \ldots, g;\ i \neq j$). Also, the estimate of $\boldsymbol{\mu}_i + \boldsymbol{\gamma}_k$ is unique and is given by $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) + \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*) = \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..} + \bar{\mathbf{y}}_{.k}$.

The ML estimate of $\boldsymbol{\Sigma}$ is equal to

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{gK}\left[\sum_{i=1}^{g}\sum_{k=1}^{K}(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*))(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*))'\right].$$

Once we substitute the estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$, we have that $\hat{\boldsymbol{\Sigma}} = \left[\frac{g-1}{g}\right]^2 \hat{\boldsymbol{\Sigma}}_D$, where

$$\hat{\boldsymbol{\Sigma}}_D = \left[\frac{1}{gK}\right]\left[\sum_{i=1}^{g}\sum_{k=1}^{K}\left(\mathbf{D}_{ik,y} - \bar{\mathbf{D}}_{i.,y}\right)\left(\mathbf{D}_{ik,y} - \bar{\mathbf{D}}_{i.,y}\right)'\right],$$

$\mathbf{D}_{ik,y} = \mathbf{y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g} \mathbf{y}_{lk}$ and $\bar{\mathbf{D}}_{i.,y} = \bar{\mathbf{y}}_{i.} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g} \bar{\mathbf{y}}_{l.}$. As in the paired case, our estimate of $\boldsymbol{\Sigma}$ is unique.

Even though $\boldsymbol{\mu}_i$ and $\boldsymbol{\gamma}_k$ are not identifiable parameters in our model, the classification regions in (3.26) can be shown to remain invariant when we apply them to the training data. When applied to the training feature measurements, the rule in (3.26) is given by

$$R_i: \left[\hat{\tilde{\mathbf{y}}}_{ik} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) + \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*))\right]' \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*)) > 0 \; j = 1, \ldots, g; \; j \neq i, \tag{3.29}$$

where $\hat{\tilde{\mathbf{y}}}_{ik} = \mathbf{y}_{ik} - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$ denotes the training feature data that have been adjusted for the effect of matching. When we plug in the formulas for $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)$, $\hat{\boldsymbol{\mu}}_j(\mathbf{c}^*)$, $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$, and $\hat{\boldsymbol{\Sigma}}$, the rule in (3.29) simplifies to

$$R_i: \left[\mathbf{D}_{ik,y} - \frac{g-2}{g-1}\left(\frac{\bar{\mathbf{y}}_{i.} + \bar{\mathbf{y}}_{j.}}{2} - \frac{1}{g-2}\sum_{\substack{l=1 \\ l\neq i,j}}^{g} \bar{\mathbf{y}}_{l.}\right)\right]' \hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{j.}) > 0, \quad (i, j = 1, \ldots, g; j \neq i) \tag{3.30}$$

after some simplification, which could also have been obtained by using the unbiased estimate $\hat{\boldsymbol{\Sigma}}_D^* = \frac{gK}{g(K-1)}\hat{\boldsymbol{\Sigma}}_D$. Since the estimated discriminant coefficient vector $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{j.})$ is unique, the fact that $\boldsymbol{\mu}_i$ is not identifiable is not problematic if we want to determine the feature variables that best discriminate between groups i and j, once we account for the effect of group matching. With regards to the Konopaske biomarker data, the estimated coefficient vectors $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{2.})$, $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{1.} - \bar{\mathbf{y}}_{3.})$, and $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{2.} - \bar{\mathbf{y}}_{3.})$ help us identify the biomarkers, once they've been adjusted for the effect of matching, that best discriminate between the haloperidol and olanzapine treatment groups, the haloperidol and sham treatment groups, and the olanzapine and sham treatment groups, respectively.

Using the adjusted training feature data $\tilde{\mathbf{y}}_{ik}$, we can estimate the probability of misclassification for the rule in (3.26) using resubstitution or K-fold cross validation (where the adjusted training feature measurements $\tilde{\mathbf{y}}_{1k}, \ldots, \tilde{\mathbf{y}}_{gk}$ for each of the K matches are omitted at a time) as described in Section 3.2.3.

If we want to use the linear discriminant rule in (3.26) to classify an individual in a new match beyond the training data, the parameter $\boldsymbol{\gamma}$ must be re-estimated for this match. As we did in the paired case, we extend Lachenbruch and Tu et al.'s conditional model in this instance to include parameters that must be re-estimated for each new match, i.e., $\boldsymbol{\gamma}$. To

elaborate, suppose we know the feature measurements for an individual and their $g-1$ siblings in a match, namely, $\mathbf{y}_{ind}$, $\mathbf{y}_{sib,1}$, $\ldots$, $\mathbf{y}_{sib,g-1}$. If we view the estimates $\hat{\boldsymbol{\mu}}_1(\mathbf{c}^*), \ldots, \hat{\boldsymbol{\mu}}_g(\mathbf{c}^*)$, and $\hat{\boldsymbol{\Sigma}}$ from the training data as fixed, we can apply our model in Section 3.5.1.1 to the random feature vectors $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ so that for a given match, $\mathbf{Y}_{ind} \sim N_P(\hat{\boldsymbol{\mu}}_{i_1}(\mathbf{c}^*)+\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$ in group i_1, $\mathbf{Y}_{sib,1} \sim N_P(\hat{\boldsymbol{\mu}}_{i_2}(\mathbf{c}^*) + \boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$ in group i_2, $\ldots$, $\mathbf{Y}_{sib,g-1} \sim N_P(\hat{\boldsymbol{\mu}}_{i_g}(\mathbf{c}^*) + \boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$ in group i_g $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g)$, where we assume $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ are mutually independent. Alternatively, we can examine $\mathbf{Y}_{ind} - \hat{\boldsymbol{\mu}}_{i_1}(\mathbf{c}^*) \equiv \mathbf{Y}^*_{ind} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$, $\mathbf{Y}_{sib,1} - \hat{\boldsymbol{\mu}}_{i_2}(\mathbf{c}^*) \equiv \mathbf{Y}^*_{sib,1} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$, $\ldots$, $\mathbf{Y}_{sib,g-1} - \hat{\boldsymbol{\mu}}_{i_g}(\mathbf{c}^*) \equiv \mathbf{Y}^*_{sib,g-1} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}})$. Using the likelihood function based on $\mathbf{y}^*_{ind}$, $\mathbf{y}^*_{sib,1}$, $\ldots$, $\mathbf{y}^*_{sib,g-1}$, the ML estimate of $\boldsymbol{\gamma}$ is given by $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*) = \frac{1}{g}(\mathbf{y}^*_{ind} + \sum_{m=1}^{g-1} \mathbf{y}^*_{sib,m}) = \frac{1}{g}\left[(\mathbf{y}_{ind} + \sum_{m=1}^{g-1} \mathbf{y}_{sib,m}) - \sum_{i=1}^{g} \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)\right]$, assuming the estimates from the training data are given. Due to the fact that $\mathbf{Y}^*_{ind}$, $\mathbf{Y}^*_{sib,1}$, $\ldots$, $\mathbf{Y}^*_{sib,g-1}$ are all identically distributed, the likelihood function remains invariant no matter which groups $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ belong to, so that the estimate $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$ is unique. Even if $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ are not mutually independent, the likelihood function for $\mathbf{y}^*_{ind}$, $\mathbf{y}^*_{sib,1}$, $\ldots$, $\mathbf{y}^*_{sib,g-1}$ remains invariant as long as the covariance matrix between any two of the g feature vectors, $\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$, remains the same and is symmetric. Once we plug the estimates $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$, $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)$, $\hat{\boldsymbol{\mu}}_j(\mathbf{c}^*)$ $(i, j = 1, \ldots, g;\ i \neq j)$, and $\hat{\boldsymbol{\Sigma}}$ into (3.26) and simplify our result, we obtain the classification regions

$$R_i: \left[\mathbf{y}_{\text{diff}} - \frac{g-2}{g-1}\left(\frac{\bar{\mathbf{y}}_{i\cdot} + \bar{\mathbf{y}}_{j\cdot}}{2} - \frac{1}{g-2}\sum_{\substack{l=1 \\ l\neq i,j}}^{g} \bar{\mathbf{y}}_{l\cdot}\right)\right]' \hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{j\cdot}) > 0, \quad j = 1, \ldots, g;\ j \neq i, \tag{3.31}$$

where $\mathbf{y}_{\text{diff}} = \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{y}_{sib,m}$. We see that the rule in (3.31) is identical to the rule in (3.28) based on the feature difference $\mathbf{y}_{\text{diff}} = \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{y}_{sib,m}$, with $\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ in (3.28) replaced with the estimate $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{i\cdot} - \bar{\mathbf{y}}_{j\cdot})$.

When we examine our estimate $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$, we see that we could not have estimated $\boldsymbol{\gamma}$ for an individual in a new match in the manner we just described if we only knew the feature data for this individual. Without an estimate of $\boldsymbol{\gamma}$, we cannot use the classification rule in (3.26) to classify this individual.

3.5.2.2 Classifying Each Member of a Given Match using Feature Difference

We can also consider an estimation approach based on our formulated model for the differ-

enced random feature vector $\mathbf{Y}_{\text{diff}} \equiv \mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}$ in Section 3.5.1.2. The parameters $\boldsymbol{\mu}_i^{\text{diff}} = \boldsymbol{\mu}_i - \sum_{\substack{l=1 \\ l\neq i}}^{g}\boldsymbol{\mu}_l$ $(i = 1, \ldots, g)$ and $\boldsymbol{\Sigma}_*$ in this model can be estimated via ML estimation using the training feature differences $\mathbf{D}_{ik,y} = \mathbf{y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g}\mathbf{y}_{lk}$, where $\mathbf{D}_{ik,y}$ is computed for the i^{th} group member of the k^{th} match in the training data $(i = 1, \ldots, g;\ k = 1, \ldots, K)$. Based on our model for $\mathbf{Y}_{\text{diff}}$, the training feature differences $\mathbf{D}_{1k,y}$, $\mathbf{D}_{2k,y}$, $\ldots$, and $\mathbf{D}_{gk,y}$ belong to the 1^{st}, 2^{nd}, $\ldots$, and g^{th} populations, respectively.

The ML estimates of $\boldsymbol{\mu}_i^{\text{diff}}$ and $\boldsymbol{\Sigma}_*$ can be shown to equal, respectively, $\bar{\mathbf{D}}_{i.,y} = \bar{\mathbf{y}}_{i.} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g}\bar{\mathbf{y}}_{l.}$ and $\hat{\boldsymbol{\Sigma}}_D$, which is defined as in Section 3.5.2.1. After we plug these estimates into (3.27), we obtain the same rule as in (3.31). In addition, when we apply the rule in (3.27) to the differenced training feature data $\mathbf{D}_{ik,y}$, we get the same estimated rule as in (3.30).

Based on the training feature differences $\mathbf{D}_{ik,y}$, we can use resubstitution or K-fold cross validation (where the differences $\mathbf{D}_{1k,y}, \ldots, \mathbf{D}_{gk,y}$ for each of the K matches in the training data are omitted at a time) as described in Section 3.1.2 to estimate the probability of misclassification for the rule in (3.27).

In examining the results we obtain from our estimation approaches in Sections 3.5.2.1 and 3.5.2.2, we see that the rule in (3.26) based on the feature vector $\tilde{\mathbf{y}}_{ind}$ that is adjusted for the effect of group matching is the same as the rule in (3.28) based on the feature difference $\mathbf{y}_{\text{diff}} = \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{y}_{sib,m}$ when applied to matched data. As a result, not only do both estimation approaches produce the same classification results in the data setting, but they also help us identify, via the estimated discriminant coefficient vector $\hat{\boldsymbol{\Sigma}}_D^{-1}(\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{j.})$, the same set of feature variables that best distinguishes group i from group j $(i, j = 1, \ldots, g;\ j \neq i)$, once we account for the effect of group matching. For example, referring back to the Konopaske biomarker data, we could apply either estimation approach to identify among the examined biomarkers a unique set that best discriminates between the haloperidol and olanzapine treatment groups, a unique set that best discriminates between the haloperidol and sham treatment groups, and a unique set that best discriminates between the olanzapine and sham treatment groups, while, at the same time, accounting for the effect of triad matching on these biomarkers.

3.5.2.3 Classifying All Members of a Given Match, with Unknown Match Effect Unlike the paired case, the differenced and stacked approaches we develop in Sections

3.5.1.2 and 3.5.1.3, respectively, to handle matching across multiple groups do not yield the same linear discriminant classification rule. Therefore, when we implement both of these approaches on actual data, we still do not obtain the same linear discriminant rule. The details on how to implement our stacked approach using the available training data are provided in Appendix B.2, where, for notational simplicity, we focus on the case where individuals are matched across three groups.

3.6 ACCOUNTING FOR EFFECTS OF MULTIPLE GROUP MATCHING AND COVARIATES IN LDA

3.6.1 Normal Populations with Known Parameters

In Section 3.6, we extend our methodology to account not only for the effects of group matching on the feature data, but also for the effects of additional covariates. For the reader, this section extends Section 3.5 in the same way Section 3.4 extended Section 3.3. and, thus, can be safely skipped.

We begin with an extension of the matched adjustment methodologies we develop in Sections 3.5.1.1, 3.5.1.2, and 3.5.1.3 to also take into account covariate effects, retaining our assumptions of equal priors and equal misclassification costs. As we noted in the paired case, we could alternately ignore the effect of group matching on the feature data and apply general covariance adjusted LDA to instead account for the effects of the variables on which individuals were matched and, thus, all comments made in the paired case are relevant here.

3.6.1.1 Classifying One Member of a Match using Known Match and Covariate Effects First, we let $(\mathbf{Y}_{ind}, \mathbf{X}_{ind})$ denote the random feature and covariate vectors for any individual in a match. For known $\boldsymbol{\gamma}$ and given $\mathbf{X}_{ind} = \mathbf{x}_{ind}$, we assume $\mathbf{Y}_{ind} \sim N_P(\boldsymbol{\mu}_i + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{ind}, \boldsymbol{\Sigma}_{(x)})$ in the i^{th} group $(i = 1, \ldots, g)$, where $\boldsymbol{\beta}$ is defined as in Section 3.4.1.1 and $\boldsymbol{\Sigma}_{(x)}$ is assumed to be common to all g groups, is independent of the value of $\mathbf{x}_{ind}$, and is assumed to be positive definite. Based on these conditional densities for $\mathbf{Y}_{ind}$, we can apply

the conditional rule in (3.7) to obtain the classification rule

$$R_{i(x)}: \left[\tilde{\mathbf{y}}_{ind(x)} - \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)\right]' \boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) > 0, \quad j = 1, \ldots, g;\ j \neq i, \tag{3.32}$$

where $\tilde{\mathbf{y}}_{ind(x)} = \mathbf{y}_{ind} - \boldsymbol{\gamma} - \boldsymbol{\beta}\mathbf{x}_{ind}$ denotes the feature measurement for an individual that has been adjusted for both matching and covariate effects. Using the rule in (3.32), we classify an individual with adjusted feature measurement $\tilde{\mathbf{y}}_{ind(x)}$ into the i^{th} group if this value falls into region $R_{i(x)}$.

Once its coefficients have been suitably standardized, we can use the adjusted linear discriminant function $\tilde{\mathbf{y}}'_{ind(x)}\boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ in (3.32) to identify the feature variables that best discriminate between groups i and j, after we adjust these feature variables for matching and covariate effects. We can also use the sign of the p^{th} $(p = 1, \ldots, P)$ element of the adjusted discriminant coefficient vector $\boldsymbol{\Sigma}_{(x)}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ to determine whether the p^{th} adjusted feature variable is larger or smaller for group i compared with group j, holding all other adjusted feature variables fixed.

As in the paired case, we can generalize our conditional model for $\mathbf{Y}_{ind}$ such that for known $\boldsymbol{\gamma}$ and given $\mathbf{X}_{ind} = \mathbf{x}_{ind}$, $\mathbf{Y}_{ind} \sim N_P(\boldsymbol{\mu}_i + \boldsymbol{\gamma} + \boldsymbol{\rho}(\mathbf{x}_{ind}; \boldsymbol{\Theta}), \boldsymbol{\Sigma}_{(x)})$ in the i^{th} group $(i = 1, \ldots, g)$, where $\boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$ is defined as in Section 3.2.3.

3.6.1.2 Classifying One Member of a Match using Covariate Adjusted Feature Difference

An alternative to the previous approach is to extend our differencing approach in Section 3.5.1.2 to also account for covariate effects on the feature data, which we can carry out by applying Lachenbruch and Tu et al.'s conditional model to the random feature difference vector $\mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}$.

First, we let $\mathbf{X}_{ind}, \mathbf{X}_{sib,1}, \ldots, \mathbf{X}_{sib,g-1}$ denote the random covariate vectors for an individual and their $g-1$ siblings in a given match, and let $\mathbf{X}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{X}_{sib,m} \equiv \mathbf{X}_{\text{diff}}$ denote the random differenced covariate vector for an individual in this match. Given $\mathbf{X}_{\text{diff}} = \mathbf{x}_{\text{diff}}$, we assume $(\mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}) \equiv \mathbf{Y}_{\text{diff}} \sim N_P(\boldsymbol{\mu}_i^{\text{diff}} + \boldsymbol{\beta}\mathbf{x}_{\text{diff}}, \boldsymbol{\Sigma}_{*(x)})$ in the i^{th} population $(i = 1, \ldots, g)$, where $\boldsymbol{\mu}_i^{\text{diff}}$ are defined as in Section 3.5.1.2. We have that $\boldsymbol{\Sigma}_{*(x)}$ remains the same in each population if the g conditional variance-covariance matrices for $\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$, along with the covariances for all pairs of the g feature vectors

$\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$ for given $\mathbf{x}_{ind}$, $\mathbf{x}_{sib,1}$, $\ldots$, $\mathbf{x}_{sib,g-1}$, do not change depending on population. As in Section 3.5.1.2, the covariance matrix between any pair of these g feature vectors need not be symmetric in order for $\boldsymbol{\Sigma}_{*(x)}$ to remain the same in each population.

Retaining our assumption from Section 3.5.1.2 that the prior probability of each population is $1/g$, we can apply general covariance adjusted LDA based on the conditional densities of $\mathbf{Y}_{\text{diff}}$ to obtain the linear discriminant rule

$$R^{\text{diff}}_{i(x)} : \left[\tilde{\mathbf{y}}_{\text{diff(x)}} - \frac{1}{2}\left(\boldsymbol{\mu}^{\text{diff}}_i + \boldsymbol{\mu}^{\text{diff}}_j\right)\right]' \boldsymbol{\Sigma}^{-1}_{*(x)}\left(\boldsymbol{\mu}^{\text{diff}}_i - \boldsymbol{\mu}^{\text{diff}}_j\right) > 0, \quad j = 1, \ldots, g;\ j \neq i, \tag{3.33}$$

where $\tilde{\mathbf{y}}_{\text{diff(x)}} = (\mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{y}_{sib,m}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{x}_{sib,m})$ is the covariate adjusted feature difference. When we plug in the formula for $\boldsymbol{\mu}^{\text{diff}}_i$, the rule in (3.33) can be re-expressed as

$$R^{\text{diff}}_{i(x)} : \left[\tilde{\mathbf{y}}_{\text{diff(x)}} - \frac{g-2}{g-1}\left(\frac{\boldsymbol{\mu}_i + \boldsymbol{\mu}_j}{2} - \frac{1}{g-2}\sum_{\substack{l=1 \\ l\neq i,j}}^{g}\boldsymbol{\mu}_l\right)\right]' \boldsymbol{\Sigma}^{-1}_{*(x)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) > 0, \quad j = 1, \ldots, g; j \neq i. \tag{3.34}$$

For each individual in a match, we compute the covariate adjusted difference $\tilde{\mathbf{y}}_{\text{diff(x)}}$ and classify this individual into the i^{th} group if this difference falls into region $R^{\text{diff}}_{i(x)}$ $(i = 1, \ldots, g)$.

From (3.34), we can use the adjusted discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}_{*(x)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ to determine which of the feature variables of interest best discriminate between groups i and j, once the effects of group matching and additional covariates on these feature variables have been accounted for.

3.6.1.3 Classifying All Members of a Match using Known Match and Covariate Effects

We can also apply the adjustment methodology we develop in Section 3.6.1.1 to the stacked feature vector $\mathbf{Y}^+$. For notational simplicity, we again focus on the case where individuals are matched across three groups. In our discussion, we define $\mathbf{Y}^+$ as in Section 3.5.1.3 and let $\mathbf{X}^+ = \begin{bmatrix}\mathbf{x}_{ind}\\ \mathbf{x}_{sib,1}\\ \mathbf{x}_{sib,2}\end{bmatrix}$ denote the random covariate vector that corresponds to an individual and their two siblings in a triad. Given $\mathbf{X}^+ = \mathbf{x}^+$, we assume $\mathbf{Y}^+ \sim N_{3P}(\boldsymbol{\mu}^+_{i(x)}, \boldsymbol{\Sigma}^+_{(x)})$ in the i^{th} group ordering $(i = 1, \ldots, 6)$ for a given match, where

$$\boldsymbol{\mu}^+_{1(x)} = \begin{bmatrix}\boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1}\\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2}\end{bmatrix},\ \boldsymbol{\mu}^+_{2(x)} = \begin{bmatrix}\boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1}\\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2}\end{bmatrix},\ \boldsymbol{\mu}^+_{3(x)} = \begin{bmatrix}\boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1}\\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2}\end{bmatrix},\ \boldsymbol{\mu}^+_{4(x)} = \begin{bmatrix}\boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1}\\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2}\end{bmatrix},$$

$$\boldsymbol{\mu}^+_{5(x)} = \begin{bmatrix}\boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1}\\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2}\end{bmatrix},\ \boldsymbol{\mu}^+_{6(x)} = \begin{bmatrix}\boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind}\\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1}\\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2}\end{bmatrix},\ \boldsymbol{\Sigma}^+_{(x)} = \begin{bmatrix}\boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Psi}\\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi} & \boldsymbol{\Psi}\\ \boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi}\end{bmatrix},$$

and $\mathbf{\Psi}$ is defined as in Section 3.5.1.3. From this conditional model for $\mathbf{Y}^+$, we can obtain a set of classification regions that allows us to simultaneously classify an individual in a triad and their two siblings in that triad into one of the six group orderings using the stacked observed feature and covariate values $\mathbf{y}^+ = \begin{bmatrix} \mathbf{y}_{ind} \\ \mathbf{y}_{sib,1} \\ \mathbf{y}_{sib,2} \end{bmatrix}$ and $\mathbf{x}^+ = \begin{bmatrix} \mathbf{x}_{ind} \\ \mathbf{x}_{sib,1} \\ \mathbf{x}_{sib,2} \end{bmatrix}$. The details for constructing the classification rule based on the conditional model for $\mathbf{Y}^+$ are provided in Appendix B.3.

When we accounted for pairing and covariate effects on the feature data, we showed that the differencing and stacked approaches in Sections 3.4.1.2 and 3.4.1.3, respectively, produced the same linear discriminant classification rule. As a result, both approaches not only yield the same classification results, but also provide us with the same information with regards to which feature variables best discriminate between the two groups under consideration, once we account for both pairing and covariate effects. However, when we proceed to the multiple group case, we have that the linear discriminant rule we obtain in Section 3.6.1.2 based on the conditional model for $\mathbf{Y}_{\text{diff}} \equiv \mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{Y}_{sib,m}$ is entirely different from the rule we obtain in Section 3.6.1.3 based on the conditional model for $\mathbf{Y}^+$. In fact, it can be shown that not only do these two approaches produce different classification results, but they also do not provide us with the same information from a discriminatory standpoint, as we saw in Section 3.5.1.3 when we only accounted for the effect of group matching on the feature data.

3.6.2 Normal Populations with Unknown Parameters

We now discuss how the adjustment procedures we develop in Sections 3.6.1.1 to 3.6.1.3 can be implemented using training data consisting of $(\mathbf{y}_{ik}, \mathbf{x}_{ik})$, the observed feature and covariate vectors for the member of the k^{th} match belonging to the i^{th} group ($i = 1, \ldots, g; k = 1, \ldots, K$).

3.6.2.1 Classifying Each Member of a Given Match, with Unknown Match and Covariate Effects Based on our conditional model in Section 3.6.1.1, we let the random feature vector $\mathbf{Y}_{ik}$, which is as defined as in Section 3.5.2.1, have conditional mean $E[\mathbf{Y}_{ik}|\mathbf{x}_{ik}] = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_k + \boldsymbol{\beta}\mathbf{x}_{ik}$ and variance-covariance matrix $\mathbf{\Sigma}_{(x)}$. Using ML estimation

to fit this model based on the training data, we assume that the design matrix for this model satisfies suitable conditions so that the ML estimate $\hat{\boldsymbol{\beta}}$ is unique. The family of estimates for $\boldsymbol{\mu}_i$ and $\boldsymbol{\gamma}_k$ are given by $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) = \bar{\mathbf{y}}_{i.} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{i.} - (\bar{\mathbf{y}}_{..} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{..}) - \mathbf{c}^*_{\mathbf{x}}$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}})$ $= \bar{\mathbf{y}}_{.k} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{.k} + \mathbf{c}^*_{\mathbf{x}}$, where $\bar{\mathbf{y}}_{i.}$, $\bar{\mathbf{y}}_{.k}$, and $\bar{\mathbf{y}}_{..}$ are defined as in Section 3.5.2.1, $\bar{\mathbf{x}}_{i.} = \frac{\sum_{k=1}^{K}\mathbf{x}_{ik}}{K}$, $\bar{\mathbf{x}}_{.k} = \frac{\sum_{i=1}^{g}\mathbf{x}_{ik}}{g}$, $\bar{\mathbf{x}}_{..} = \frac{\sum_{i=1}^{g}\sum_{k=1}^{K}\mathbf{x}_{ik}}{gK}$, and $\mathbf{c}^*_{\mathbf{x}} = -(\bar{\mathbf{y}}_{..} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{..}) + \mathbf{c}$. We then have that the ML estimate of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$ is given by $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*) - \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*) = \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{j.} - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{i.} - \bar{\mathbf{x}}_{j.})$ $(i, j = 1, \ldots, g;\, i \neq j)$.

The ML estimate of $\boldsymbol{\Sigma}_{(x)}$ is equal to

$$\hat{\boldsymbol{\Sigma}}_{(x)} = \frac{1}{gK}\left[\sum_{i=1}^{g}\sum_{k=1}^{K}(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik})(\mathbf{y}_{ik} - \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik})'\right].$$

After we substitute the estimates $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*)$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$, we have that $\hat{\boldsymbol{\Sigma}}_{(x)} = \left[\frac{g-1}{g}\right]^2 \hat{\boldsymbol{\Sigma}}_{D(x)}$, where

$$\hat{\boldsymbol{\Sigma}}_{D(x)} = \frac{1}{gK}\left[\sum_{i=1}^{g}\sum_{k=1}^{K}\left(\hat{\mathbf{D}}^{adj}_{ik,y} - \bar{\hat{\mathbf{D}}}^{adj}_{i.,y}\right)\left(\hat{\mathbf{D}}_{ik,y} - \bar{\hat{\mathbf{D}}}^{adj}_{i.,y}\right)'\right],$$

$\hat{\mathbf{D}}^{adj}_{ik,y} = \mathbf{D}_{ik,y} - \hat{\boldsymbol{\beta}}\mathbf{D}_{ik,x} = (\mathbf{y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1\\l\neq i}}^{g}\mathbf{y}_{lk}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1\\l\neq i}}^{g}\mathbf{x}_{lk})$, and $\bar{\hat{\mathbf{D}}}^{adj}_{i.,y} = \bar{\mathbf{D}}_{i.,y} - \hat{\boldsymbol{\beta}}\bar{\mathbf{D}}_{i.,x}$ $= (\bar{\mathbf{y}}_{i.} - \frac{1}{g-1}\sum_{\substack{l=1\\l\neq i}}^{g}\bar{\mathbf{y}}_{l.}) - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{i.} - \frac{1}{g-1}\sum_{\substack{l=1\\l\neq i}}^{g}\bar{\mathbf{x}}_{l.})$.

When applied to the training data, the rule in (3.32) has the form

$$R_{i(x)} : \left[\hat{\tilde{\mathbf{y}}}_{ik(x)} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) + \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*_{\mathbf{x}}))\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{(x)}(\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\mu}}_j(\mathbf{c}^*_{\mathbf{x}})) > 0, \;\; j = 1, \ldots, g;\, j \neq i, \quad (3.35)$$

where $\hat{\tilde{\mathbf{y}}}_{ik(x)} = \mathbf{y}_{ik} - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik}$ denotes the training feature data that have been adjusted for matching and covariate effects. The estimated rule in (3.35) can also be expressed as

$$R_{i(x)} : \left[\hat{\mathbf{D}}^{adj}_{ik,y} - \frac{g-2}{g-1}\left(\frac{\bar{\mathbf{y}}^*_{i.} + \bar{\mathbf{y}}^*_{j.}}{2} - \frac{1}{g-2}\sum_{\substack{l=1\\l\neq i,j}}^{g}\bar{\mathbf{y}}^*_{l.}\right)\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{D(x)}(\bar{\mathbf{y}}^*_{i.} - \bar{\mathbf{y}}^*_{j.}) > 0, \; j = 1, \ldots, g; j \neq i, \quad (3.36)$$

where $\bar{\mathbf{y}}^*_{i.} = \bar{\mathbf{y}}_{i.} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{i.}$, $\bar{\mathbf{y}}^*_{j.} = \bar{\mathbf{y}}_{j.} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{j.}$, and $\bar{\mathbf{y}}^*_{l.} = \bar{\mathbf{y}}_{l.} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{l.}$. We could also have obtained (3.36) by using the unbiased estimate $\hat{\boldsymbol{\Sigma}}^*_{D(x)} = \frac{gK}{g(K-1)}\hat{\boldsymbol{\Sigma}}_{D(x)}$. Based on the adjusted training feature data $\tilde{\mathbf{y}}_{ik(x)}$, we can use resubstitution or K-fold cross validation as described in Section 3.5.2.1 to estimate the probability of misclassification for the conditional rule in (3.32).

To use the rule in (3.32) to classify an individual in a match that is not part of the training data, we must re-estimate $\boldsymbol{\gamma}$ for this match. Suppose we are provided with the

feature and covariate measurements for this individual and their $g-1$ siblings in that match, i.e., $(\mathbf{y}_{ind}, \mathbf{x}_{ind})$, $(\mathbf{y}_{sib,1}, \mathbf{x}_{sib,1})$, $\ldots$, $(\mathbf{y}_{sib,g-1}, \mathbf{x}_{sib,g-1})$. Using the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$ $(i = 1, \ldots, g)$, and $\hat{\boldsymbol{\Sigma}}_{(x)}$ from the training data, we assume for a given match and given $\mathbf{x}_{ind}$, $\mathbf{x}_{sib,1}$, $\ldots$, $\mathbf{x}_{sib,g-1}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$, and $\hat{\boldsymbol{\Sigma}}_{(x)}$, $\mathbf{Y}_{ind} \sim N_P(\hat{\boldsymbol{\mu}}_{i_1}(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{ind}, \hat{\boldsymbol{\Sigma}}_{(x)})$ in group i_1, $\mathbf{Y}_{sib,1} \sim N_P(\hat{\boldsymbol{\mu}}_{i_2}(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,1}, \hat{\boldsymbol{\Sigma}}_{(x)})$ in group i_2, $\ldots$, $\mathbf{Y}_{sib,g-1} \sim N_P(\hat{\boldsymbol{\mu}}_{i_g}(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,g-1}, \hat{\boldsymbol{\Sigma}}_{(x)})$ in group i_g $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g)$, where we again assume $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ are mutually independent. We can also consider $\mathbf{Y}_{ind} - \hat{\boldsymbol{\mu}}_{i_1}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ind} \equiv \mathbf{Y}^{*(x)}_{ind} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}}_{(x)})$, $\mathbf{Y}_{sib,1} - \hat{\boldsymbol{\mu}}_{i_2}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,1} \equiv \mathbf{Y}^{*(x)}_{sib,1} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}}_{(x)})$, $\ldots$, $\mathbf{Y}_{sib,g-1} - \hat{\boldsymbol{\mu}}_{i_g}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,g-1} \equiv \mathbf{Y}^{*(x)}_{sib,g-1} \sim N_P(\boldsymbol{\gamma}, \hat{\boldsymbol{\Sigma}}_{(x)})$. From the likelihood function based on $\mathbf{Y}^{*(x)}_{ind}, \mathbf{Y}^{*(x)}_{sib,1}, \ldots, \mathbf{Y}^{*(x)}_{sib,g-1}$, we obtain the ML estimate $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}}) = \frac{1}{g}(\mathbf{y}^{*(x)}_{ind} + \sum_{m=1}^{g-1} \mathbf{y}^{*(x)}_{sib,m}) = \frac{1}{g}\left[(\mathbf{y}_{ind} + \sum_{m=1}^{g-1} \mathbf{y}_{sib,m}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} + \sum_{m=1}^{g-1} \mathbf{x}_{sib,m})\right] - \frac{1}{g}\sum_{i=1}^{g} \hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$. We have that $\mathbf{Y}^{*(x)}_{ind}$, $\mathbf{Y}^{*(x)}_{sib,1}$, $\ldots$, $\mathbf{Y}^{*(x)}_{sib,g-1}$ are identically distributed, so that the corresponding likelihood function remains invariant no matter which groups $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ belong to, which implies that $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}})$ is unique. Even if $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ are not mutually independent, but the covariance matrix between any two of these g vectors remains the same and is symmetric, the invariance of the likelihood function based on $\mathbf{y}^{*(x)}_{ind}$, $\mathbf{y}^{*(x)}_{sib,1}$, $\ldots$, $\mathbf{y}^{*(x)}_{sib,g-1}$ would still hold. When we substitute the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_i(\mathbf{c}^*_{\mathbf{x}})$, $\hat{\boldsymbol{\mu}}_j(\mathbf{c}^*_{\mathbf{x}})$ $(i, j = 1, \ldots, g;\ i \neq j)$, $\hat{\boldsymbol{\Sigma}}_{(x)}$, and $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}})$, the rule in (3.32) takes on the form

$$R_{i(x)}: \left[\hat{\tilde{\mathbf{y}}}_{\text{diff(x)}} - \frac{g-2}{g-1}\left(\frac{\bar{\mathbf{y}}^*_{i.} + \bar{\mathbf{y}}^*_{j.}}{2} - \frac{1}{g-2}\sum_{\substack{l=1 \\ l\neq i,j}}^{g} \bar{\mathbf{y}}^*_{l.}\right)\right]' \hat{\boldsymbol{\Sigma}}^{-1}_{D(x)}(\bar{\mathbf{y}}^*_{i.} - \bar{\mathbf{y}}^*_{j.}) > 0, \quad j = 1, \ldots, g; j \neq i, \tag{3.37}$$

after some simplification, where $\hat{\tilde{\mathbf{y}}}_{\text{diff(x)}} = (\mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{y}_{sib,m}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{x}_{sib,m})$. We have that the rule in (3.37) is identical to the rule in (3.34), with $\boldsymbol{\beta}$ and the adjusted discriminant coefficient vector $\boldsymbol{\Sigma}^{-1}_{*(x)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ in (3.34) replaced with the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}^{-1}_{D(x)}(\bar{\mathbf{y}}^*_{i.} - \bar{\mathbf{y}}^*_{j.})$, respectively.

3.6.2.2 Classifying Each Member of a Given Match using Covariate Adjusted Feature Difference An alternate estimation approach we can take is to implement the differencing approach we develop in Section 3.6.1.2 using the available training data. We begin by letting $\mathbf{D}_{ik,Y} \equiv \mathbf{Y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g} \mathbf{Y}_{lk}$ denote the random differenced feature vector

for the member of the k^{th} match belonging to the i^{th} group ($i = 1, \ldots, g$; $k = 1, \ldots, K$), i.e., in the k^{th} match, $\mathbf{D}_{ik,Y}$ corresponds to the i^{th} population. Based on our conditional model for the feature difference $\mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}$ in Section 3.6.1.2, we assume $\mathbf{D}_{ik,Y}$ has conditional mean $E[\mathbf{D}_{ik,Y}|\mathbf{D}_{ik,x}] = \boldsymbol{\mu}_i^{\text{diff}} + \boldsymbol{\beta}\mathbf{D}_{ik,x}$ and conditional variance-covariance matrix $\boldsymbol{\Sigma}_{*(x)}$. We then fit the model for $E[\mathbf{D}_{ik,Y}|\mathbf{D}_{ik,x}]$ using ML estimation based on the differences $(\mathbf{D}_{ik,y}, \mathbf{D}_{ik,x})$, which are defined as in Sections and 3.5.2.1 and 3.6.2.1, and retain our assumption that the design matrix for our model satisfies suitable conditions so that the ML estimate $\hat{\boldsymbol{\beta}}$ is unique. In fitting this model, we obtain the estimates $\hat{\boldsymbol{\mu}}_i^{\text{diff}} = \bar{\hat{\mathbf{D}}}_{i.,y}^{adj}$ ($i = 1, \ldots, g$) and $\hat{\boldsymbol{\Sigma}}_{*(x)} = \hat{\boldsymbol{\Sigma}}_{D(x)}$, where $\bar{\hat{\mathbf{D}}}_{i.,y}^{adj}$ and $\hat{\boldsymbol{\Sigma}}_{D(x)}$ are defined as in Section 3.6.2.1. When we plug the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}_i^{\text{diff}}$, $\hat{\boldsymbol{\mu}}_j^{\text{diff}}$ ($i, j = 1, \ldots, g$; $i \neq j$), and $\hat{\boldsymbol{\Sigma}}_{*(x)}$ into (3.33), we obtain the same rule as in (3.37). Also, when we apply the rule in (3.33) to the training data, we get the same estimated rule as in (3.36).

Using the covariate adjusted training feature differences $\hat{\mathbf{D}}_{ik,y}^{adj}$, which are defined as in Section 3.6.2.1, we can use resubstitution or K-fold cross validation (where the covariate adjusted differences $\hat{\mathbf{D}}_{1k,y}^{adj}$, $\ldots$, $\hat{\mathbf{D}}_{gk,y}^{adj}$ for each of the K matches in the training data are omitted at a time) as described in Section 3.2.3 to estimate the probability of misclassification associated with the conditional rules in (3.33) and, equivalently, (3.34).

Based on our results in Sections 3.6.2.1 and 3.6.2.2, we have that the rule in (3.32) based on the adjusted feature vector $\tilde{\mathbf{y}}_{ind(x)}$ is the same as the rule in (3.34) based on the covariate adjusted feature difference $\tilde{\mathbf{y}}_{\text{diff(x)}}$ when applied to matched data. Therefore, not only do the estimation approaches we develop in these two sections yield the same classification results in the data setting, but they also help us identify, via the estimated adjusted discriminant coefficient vector $\hat{\boldsymbol{\Sigma}}_{D(x)}^{-1}(\bar{\mathbf{y}}_{i.}^{*} - \bar{\mathbf{y}}_{j.}^{*})$, the same set of feature variables that best distinguishes group i from group j ($i, j = 1, \ldots, g$; $j \neq i$), once we account for the effects of both matching and additional covariates on the feature data.

3.6.2.3 Classifying All Members of a Given Match, with Unknown Match and Covariate Effects It can be shown that when implemented on matched data, the differencing and stacked approaches we develop in Sections 3.6.1.2 and 3.6.1.3 do not yield the same classification rule. We give a detailed discussion of how our stacked approach in Section 3.6.1.3 can be applied in the data setting in Appendix B.4.

4.0 ACCOUNTING FOR MATCHING AND COVARIATE EFFECTS IN CLASSIFICATION TREES

4.1 TRADITIONAL CLASSIFICATION TREES

4.1.1 Overview

In general, linear discriminant analysis uses training data to compute a linear rule that splits the feature space $\mathcal{Y}$ into g distinct subsets, such that a new subject who falls into one of these subsets is classified into one of the g groups. In the same spirit, classification trees were initially developed by various authors, including Morgan and Sonquist [25], Morgan and Messenger [24], and Friedman [10], to obtain from training data a nonlinear rule from which to classify new subjects, as well as determine the feature variables that best discriminate among the g groups under consideration. Hand [13] notes that Breiman, Friedman, Olshen, and Stone (BFOS) were the first authors to formally integrate and provide a theoretical justification for all previously developed tree construction procedures in their 1984 book [7]. In addition, the BFOS algorithm for recursive partitioning of the feature space is still among the most popular tree construction procedures. Thus, it is their work that is summarized in our subsequent discussion of classification trees.

Classification trees are constructed by using training data to recursively split on the coordinate axes until some stopping criterion is satisfied, such that the feature measurements in each resulting subset are as homogeneous as possible with respect to group. A new subject that falls into one of these subsets is then classified into the group most commonly represented in the subset. The feature variables that are chosen to split the feature space $\mathcal{Y}$ in this manner are those that best discriminate among these groups of interest, which is what interests us most in our application of classification trees, even though classification

tends to be the main goal in general. Although classification trees can also split $\mathcal{Y}$ using linear combinations of the feature variables, it is very computationally intensive to do so and may yield results that are hard to interpret. Thus, we do not discuss this approach here. Classification trees are usually constructed by using binary splits to iteratively split each subset of the feature space. This iterative process is typically described using decision theoretic tree notation, in which *node* t denotes a subset of the feature space and *root node* t_0 denotes the entire feature space. On the other hand, there is no recursive partitioning involved in traditional LDA, where a set of $\binom{g}{2}$ hyperplanes is used to split the entire feature space.

Traditional LDA is usually developed primarily from a theoretical or population based standpoint, where the resulting classification rule is usually first constructed under the assumption that the feature data come from a normal population with known parameters for each group under consideration. The parameters must then be estimated from training data. This LDA development is in distinction to the development of classification trees, which is typically data driven.

Our discussion of the construction of classification trees in the general case of g groups using the BFOS recursive partitioning algorithm is first done from a population based perspective, as is typically done in traditional LDA. We amplify the approach Friedman initially took in his development of classification trees [10][30], and that which Shang and Breiman took in their preliminary development of distribution based trees [31]. In particular, Friedman briefly discussed how, in the case of two known continuous distributions, the well known Kolmogorov-Smirnov distance measure could be used to determine the optimal set of splitting variables and cutpoints for a specific classification tree. Similarly, Shang and Breiman used the Gini index (see Section 4.1.2.2) to achieve the same goal in the case of at least two known continuous distributions. Our goal is to prepare a deeper understanding of the theoretical roots of the BFOS algorithm, so that we can introduce carefully the use of covariates in our tree construction procedure. In this chapter, we first consider classification trees using g known continuous distributions. For example, we show how such trees can be developed for normal populations. We then show how classification trees are constructed when the distribution functions for the g groups must be estimated from available training data.

4.1.2 Known Distributions

4.1.2.1 Tree Construction Procedure Let $\mathbf{Y} = (Y_1, \ldots, Y_P)'$ have prior probability π_i $(i = 1, \ldots, g)$ of belonging to group i, in which $\mathbf{Y}$ has some known continuous cumulative distribution function (CDF) $F_{\mathbf{Y}}^{(i)}(\centerdot)$. In addition, we assume equal misclassification costs.

To construct a tree, we first choose a feature variable Y, that is, one of the components of $\mathbf{Y}$, and a cutpoint c in $\mathbb{R}$ that splits the root node t_0 into descendant nodes t_L and t_R, such that a specific goodness of split (GOS) criterion (described in Section 4.1.2.2) defined by the split $Y \leq c$ is maximized. The basic idea is to choose Y and c to maximize the group purity (homogeneity) of nodes t_L and t_R. This splitting procedure is then applied recursively to t_L, t_R, and all subsequent descendant nodes until further splitting ceases to significantly increase group homogeneity, as defined by some specific criterion. Once splitting stops, we let T' denote a tree obtained in this manner. Nodes not split in T' are called terminal nodes and we let $\tilde{T}'$ denote the set of terminal nodes of T'.

Based on our assumption of equal misclassification costs, each terminal node t in $\tilde{T}'$ is assigned to group i if $P(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t) > P(\mathbf{Y} \in \text{group } j | \mathbf{Y} \in t)$ $(j = 1, \ldots, g;\ j \neq i)$. We use the fact that

$$P(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t) = \frac{\pi_i P^{(i)}(\mathbf{Y} \in t)}{P(\mathbf{Y} \in t)}, \tag{4.1}$$

where $P^{(i)}(\mathbf{Y} \in t) = P(\mathbf{Y} \in t | \mathbf{Y} \in \text{group } i)$, to re-express our group assignment rule of terminal node t as

$$R_i : \left\{ t : \frac{P^{(i)}(\mathbf{Y} \in t)}{P^{(j)}(\mathbf{Y} \in t)} > \frac{\pi_j}{\pi_i} \right\}, \quad j = 1, \ldots, g;\ j \neq i. \tag{4.2}$$

In the case of two groups, the rule in (4.2) can be expressed as

$$R_1 : \left\{ t : \frac{P^{(1)}(\mathbf{Y} \in t)}{P^{(2)}(\mathbf{Y} \in t)} \geq \frac{\pi_2}{\pi_1} \right\}, \quad R_2 : \left\{ t : \frac{P^{(1)}(\mathbf{Y} \in t)}{P^{(2)}(\mathbf{Y} \in t)} < \frac{\pi_2}{\pi_1} \right\}.$$

Once we obtain T', we can use it to classify an individual based on their observed feature vector $\mathbf{y}$.

Using the rule in (4.2), we have that the true misclassification rate for node t is equal to

$$P(\mathbf{Y} \in t) - max_{i=1,\ldots,g} \left[\pi_i P^{(i)}(\mathbf{Y} \in t) \right], \tag{4.3}$$

where

$$P(\mathbf{Y} \in t) = \sum_{i=1}^{g} \pi_i P^{(i)}(\mathbf{Y} \in t). \tag{4.4}$$

For each node t in the set of terminal nodes $\tilde{T}'$, we compute this error rate so that we may obtain the following true misclassification rate for tree T'

$$R^*(T') = \sum_{t \in \tilde{T}'} \left\{ P(\mathbf{Y} \in t) - max_{i=1,\ldots,g} \left[\pi_i P^{(i)}(\mathbf{Y} \in t) \right] \right\},$$

which can be expressed as $\sum_{t \in \tilde{T}'} min_{i=1,2} \left[\pi_i P^{(i)}(\mathbf{Y} \in t) \right]$ for two groups. BFOS prove that no other group assignment rule for a given tree T' yields a misclassification rate lower than $R^*(T')$ [7]. When π_i and $F_{\mathbf{Y}}^{(i)}(\cdot)$ are known, $R^*(T')$ can be computed.

We now discuss the two GOS criteria that are used in the BFOS recursive partitioning algorithm to determine the optimal split for a particular node, assuming equal misclassification costs. Without loss of generality, we assume these costs are equal to one.

4.1.2.2 GOS Criteria In choosing the feature variable Y and cutpoint c that best splits a particular node t into descendant nodes t_L and t_R, one GOS criterion that is commonly used in the data setting is based on a measure of impurity for node t, which is denoted $M(t)$. For g groups, $M(t) = \phi\left(P(\mathbf{Y} \in \text{group } 1 | \mathbf{Y} \in t), \ldots, P(\mathbf{Y} \in \text{group g} | \mathbf{Y} \in t)\right)$, where $\phi(\cdot)$ is a function defined on the set of g-tuples of numbers $(p_1, \ldots, p_g)$ such that $p_i \geq 0$ $(i = 1, \ldots, g)$ and $\sum_{i=1}^{g} p_i = 1$. We have here that $\phi(\cdot)$ is an impurity function of $p_1, \ldots, p_g$ since it is maximized when $p_i = \frac{1}{g}$, minimized when $p_i = 1$ and $p_j = 0$ $(j \neq i)$, and is a symmetric function of $p_1, \ldots, p_g$ [7]. This impurity measure based GOS criterion, which we rephrase for use in the case of known distributions, is

$$M(t) - [P(\mathbf{Y} \in t_L | \mathbf{Y} \in t) M(t_L) + P(\mathbf{Y} \in t_R | \mathbf{Y} \in t) M(t_R)] \tag{4.5}$$

(see [7][13][23]). If we choose Y and c to maximize (4.5) over all Y in $\mathbf{Y}$ and all $c \in \mathbb{R}$, then we can also choose Y and c to minimize the bracketed quantity in (4.5), which can be shown to equal

$$\frac{P(\mathbf{Y} \in t_L) M(t_L) + P(\mathbf{Y} \in t_R) M(t_R)}{P(\mathbf{Y} \in t)}. \tag{4.6}$$

The fact that the bracketed quantity in (4.5) is equal to the quantity in (4.6) holds due to the fact that t_L and t_R are both subsets of node t.

Another GOS criterion that has been used primarily in the data setting is the *twoing* criterion (see [5][7][13][23][30]), which can be defined from a population based standpoint as

$$\frac{1}{4\left[P(\mathbf{Y}\in t)\right]^2}\times P(\mathbf{Y}\in t_L)\times P(\mathbf{Y}\in t_R)\times \left[\sum_{i=1}^{g}\left|P(\mathbf{Y}\in \text{group } i|\mathbf{Y}\in t_L)-P(\mathbf{Y}\in \text{group } i|\mathbf{Y}\in t_R)\right|\right]^2. \tag{4.7}$$

Unlike the GOS criterion in (4.5), the twoing criterion is not dependent on a particular impurity measure $M(t)$.

However, we note that both the impurity measure based GOS criterion and the twoing criterion given by (4.5) and (4.7), respectively, are functions of probabilities of the form $P(\mathbf{Y}\in t)$ and $P(\mathbf{Y}\in \text{group } i|\mathbf{Y}\in t)$, which can be easily re-expressed using (4.1) and (4.4). Thus, for example, the impurity measure based GOS criterion is given by

$$\begin{aligned} M(t) &= \phi\left(\frac{\pi_1 P^{(1)}(\mathbf{Y}\in t)}{P(\mathbf{Y}\in t)},\ldots,\frac{\pi_g P^{(g)}(\mathbf{Y}\in t)}{P(\mathbf{Y}\in t)}\right)\\ &= \phi\left(\frac{\pi_1 P^{(1)}(\mathbf{Y}\in t)}{\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t)},\ldots,\frac{\pi_g P^{(g)}(\mathbf{Y}\in t)}{\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t)}\right). \end{aligned}$$

Similarly, the twoing criterion in (4.7) can be re-expressed as

$$\frac{1}{4\left[\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t)\right]^2}\times\left[\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t_L)\right]\times\left[\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t_R)\right]\times \left[\sum_{i=1}^{g}\pi_i\left|\frac{P^{(i)}(\mathbf{Y}\in t_L)}{\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t_L)}-\frac{P^{(i)}(\mathbf{Y}\in t_R)}{\sum_{j=1}^{g}\pi_j P^{(j)}(\mathbf{Y}\in t_R)}\right|\right]^2. \tag{4.8}$$

One particular impurity measure used in the literature is the Gini index, which is defined by $M_G(t)\equiv 1-\sum_{i=1}^{g}\left[P(\mathbf{Y}\in \text{group } i|\mathbf{Y}\in t)\right]^2$ for g groups (see [5][13][14][30][32]). Based on (4.4) and the fact that $P(\mathbf{Y}\in t)M_G(t)=2\sum_{\substack{i,j=1\\ i<j}}^{g}\pi_i\pi_j\left[\frac{P^{(i)}(\mathbf{Y}\in t)P^{(j)}(\mathbf{Y}\in t)}{\sum_{l=1}^{g}\pi_l P^{(l)}(\mathbf{Y}\in t)}\right]$ for g groups, the quantity we seek to minimize in (4.6) based on the Gini index $M_G(t)$ can be expressed as

$$\frac{2}{\sum_{l=1}^{g}\pi_l P^{(l)}(\mathbf{Y}\in t)}\sum_{\substack{i,j=1\\ i<j}}^{g}\pi_i\pi_j\left[\frac{P^{(i)}(\mathbf{Y}\in t_L)P^{(j)}(\mathbf{Y}\in t_L)}{\sum_{l=1}^{g}\pi_l P^{(l)}(\mathbf{Y}\in t_L)}+\frac{P^{(i)}(\mathbf{Y}\in t_R)P^{(j)}(\mathbf{Y}\in t_R)}{\sum_{l=1}^{g}\pi_l P^{(l)}(\mathbf{Y}\in t_R)}\right]. \tag{4.9}$$

In the case of two groups, the Gini index reduces to

$$\begin{aligned} M_G(t) &= 2P(\mathbf{Y} \in \text{group } 1|\mathbf{Y} \in t)P(\mathbf{Y} \in \text{group } 2|\mathbf{Y} \in t) \\ &= 2\pi_1\pi_2 \frac{P^{(1)}(\mathbf{Y} \in t)P^{(2)}(\mathbf{Y} \in t)}{[P(\mathbf{Y} \in t)]^2} \\ &= 2\pi_1\pi_2 \frac{P^{(1)}(\mathbf{Y} \in t)P^{(2)}(\mathbf{Y} \in t)}{\left[\sum_{j=1}^2 \pi_j P^{(j)}(\mathbf{Y} \in t)\right]^2} \end{aligned}$$

and $P(\mathbf{Y} \in t)M_G(t) = 2\pi_1\pi_2 \frac{P^{(1)}(\mathbf{Y}\in t)P^{(2)}(\mathbf{Y}\in t)}{\sum_{j=1}^2 \pi_j P^{(j)}(\mathbf{Y}\in t)}$, so that (4.9) can be expressed as

$$\frac{2\pi_1\pi_2}{\sum_{j=1}^2 \pi_j P^{(j)}(\mathbf{Y} \in t)} \left[\frac{P^{(1)}(\mathbf{Y} \in t_L)P^{(2)}(\mathbf{Y} \in t_L)}{\sum_{j=1}^2 \pi_j P^{(j)}(\mathbf{Y} \in t_L)} + \frac{P^{(1)}(\mathbf{Y} \in t_R)P^{(2)}(\mathbf{Y} \in t_R)}{\sum_{j=1}^2 \pi_j P^{(j)}(\mathbf{Y} \in t_R)} \right].$$

Another impurity measure that is commonly used in the literature is the cross-entropy or Deviance index $M_D(t)$, where

$$\begin{aligned} M_D(t) &= -\sum_{i=1}^g P(\mathbf{Y} \in \text{group } i|\mathbf{Y} \in t) \log\left[P(\mathbf{Y} \in \text{group } i|\mathbf{Y} \in t)\right] \\ &= -\sum_{i=1}^g \frac{\pi_i P^{(i)}(\mathbf{Y} \in t)}{P(\mathbf{Y} \in t)} \log\left[\frac{\pi_i P^{(i)}(\mathbf{Y} \in t)}{P(\mathbf{Y} \in t)}\right] \\ &= -\sum_{i=1}^g \frac{\pi_i P^{(i)}(\mathbf{Y} \in t)}{\sum_{j=1}^g \pi_j P^{(j)}(\mathbf{Y} \in t)} \log\left[\frac{\pi_i P^{(i)}(\mathbf{Y} \in t)}{\sum_{j=1}^g \pi_j P^{(j)}(\mathbf{Y} \in t)}\right] \end{aligned}$$

(see [5][13][14][30][32]). Based on (4.4) and the fact that $P(\mathbf{Y} \in t)M_D(t) = -\sum_{i=1}^g \text{dev}_{i,t}$, where $\text{dev}_{i,t} = \pi_i P^{(i)}(\mathbf{Y} \in t) \log\left[\frac{\pi_i P^{(i)}(\mathbf{Y}\in t)}{\sum_{j=1}^g \pi_j P^{(j)}(\mathbf{Y}\in t)}\right]$, the quantity we seek to minimize in (4.6) for the Deviance index $M_D(t)$ can be expressed as

$$-\frac{1}{\sum_{j=1}^g \pi_j P^{(j)}(\mathbf{Y} \in t)} \sum_{i=1}^g \left(\text{dev}_{i,t_L} + \text{dev}_{i,t_R}\right). \tag{4.10}$$

Although the Gini and Deviance indices are among several impurity measures that can be used in the construction of classification trees using the BFOS algorithm (see [13][14][30]), they are the ones that are most commonly used due to the fact that they are strictly concave functions of $P(\mathbf{Y} \in \text{group } i|\mathbf{Y} \in t)$ [7][14]. Since $\mathbf{Y}$ is assumed to be continuous, this property ensures that the impurity measure based GOS criterion in (4.5) is always positive, which is shown in Proposition C.1.1 in Appendix C.1. In other words, for continuous $\mathbf{Y}$, the use of a strictly concave impurity function in (4.5) ensures that the impurity of node t is always decreased when it is split.

We note one interesting fact regarding the connection of the GOS criterion in (4.5) to the majorization concept of Schur concavity. A function $\omega(\centerdot)$ has the previously noted properties of the impurity function $\phi(\centerdot)$ and is strictly concave if and only if $\omega(\centerdot)$ is strictly Schur concave, i.e., symmetric and strictly concave, of which the Gini and Deviance indices are two examples [2][21]. Thus, if the measure $M(t)$ is based on any strictly Schur concave function $\omega(\centerdot)$, the GOS criterion in (4.5) remains positive for continuous $\mathbf{Y}$.

Along with our group assignment rule for node t, we have that both GOS criteria and all impurity measures $M(t)$ can be computed using only our knowledge of π_i and $F_{\mathbf{Y}}^{(i)}(\centerdot)$ $(i = 1, \ldots, g)$, from which we can compute $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$.

4.1.2.3 Tree Construction Procedure for Known Normal Populations

It is interesting to note that if we retain our assumption from traditional LDA that $\mathbf{Y} = (Y_1, \ldots, Y_P)' \sim N_P(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\Sigma}_{YY})$ in the i^{th} group $(i = 1, \ldots, g)$, we can use the population version of the BFOS algorithm to obtain another approach for discriminating in the classical LDA normal setting. To elaborate, in the case where $g = 2$, suppose that $\pi_1 = \pi_2 = 0.5$. In the first step of our tree construction procedure, we seek to split the root node t_0 into t_L and t_R and use the fact that $\frac{Y_p - \mu_{Y,p,i}}{\sqrt{\sigma_{pp}}} \sim N(0,1)$ for the i^{th} group $(p = 1, \ldots, P)$. We then proceed to find, for each feature variable Y_p, the cutpoint $\grave{c}_{1,p}$ that minimizes the quantities in (4.9) or (4.10) or maximizes the quantity in (4.8), depending on the GOS criterion and impurity measure that is chosen. For the three quantities in (4.8), (4.9), and (4.10), we have that $P(\mathbf{Y} \in t) = P(\mathbf{Y} \in t_0) = 1$ because t_0 is the entire feature space. For each Y_p and $\grave{c}_{1,p}$, $P^{(i)}(\mathbf{Y} \in t_L) = P^{(i)}(Y_p \leq \grave{c}_{1,p}) = \Phi\left(\frac{\grave{c}_{1,p} - \mu_{Y,p,i}}{\sqrt{\sigma_{pp}}}\right)$ and $P^{(i)}(\mathbf{Y} \in t_R) = P^{(i)}(Y_p > \grave{c}_{1,p}) = \Phi\left(\frac{\mu_{Y,p,i} - \grave{c}_{1,p}}{\sqrt{\sigma_{pp}}}\right)$. Using the differentiability of the Gini and Deviance indices and the twoing criterion, we can show that $\grave{c}_{1,p} = \frac{1}{2}(\mu_{Y,p,1} + \mu_{Y,p,2})$. We then let c_ν denote the cutpoint $\grave{c}_{1,p}$ that minimizes (4.9) or (4.10) or maximizes (4.8) across all p and let Y_ν denote the feature variable that corresponds to this particular cutpoint. To gain further insight into what this particular result entails, suppose that $\mathbf{Y}$ is univariate (i.e., $P = 1$), so that $\mathbf{Y} = Y_1$. Then, the optimal cutpoint $\frac{1}{2}(\mu_{Y,1,1} + \mu_{Y,1,2})$ corresponding to Y_1 that first splits the feature space $\mathcal{Y}$ using the BFOS algorithm is the same cutpoint used to split $\mathcal{Y}$ in traditional LDA for two groups. Thus, if splitting were to terminate at this point, the true misclassification rates for our resulting tree T' and the linear discriminant rule given in (3.4) in Section 3.1.2 would be identical.

However, we note that the parallel between the results obtained from traditional LDA and the BFOS algorithm for univariate $\mathbf{Y}$ when we first split $\mathcal{Y}$ can be shown to no longer exist when we deal with more than two groups.

We now consider the case of general P. Suppose we seek to split the left daughter node t_L of t_0. It follows that $P(\mathbf{Y} \in t)$ is now equal to $P(Y_\nu \leq c_\nu) = 0.5\left[\Phi\left(\frac{c_\nu - \mu_{Y,\nu,1}}{\sqrt{\sigma_{\nu\nu}}}\right) + \Phi\left(\frac{c_\nu - \mu_{Y,\nu,2}}{\sqrt{\sigma_{\nu\nu}}}\right)\right]$. For each Y_p and cutpoint $\grave{c}_{2,p}$ $(p = 1, \ldots, P)$, we also have that $P^{(i)}(\mathbf{Y} \in t_L) = P^{(i)}(Y_\nu \leq c_\nu, Y_p \leq \grave{c}_{2,p})$ and $P^{(i)}(\mathbf{Y} \in t_R) = P^{(i)}(Y_\nu \leq c_\nu, Y_p > \grave{c}_{2,p}) = P^{(i)}(Y_\nu \leq c_\nu) - P^{(i)}(Y_\nu \leq c_\nu, Y_p \leq \grave{c}_{2,p})$, where $P^{(i)}(Y_\nu \leq c_\nu, Y_p \leq \grave{c}_{2,p})$ is the bivariate normal CDF corresponding to $(Y_\nu, Y_p)'$ in the i^{th} group $(i = 1, 2; p = 1, \ldots, P; Y_\nu \neq Y_p)$. For the case in which $Y_p = Y_\nu$, $\grave{c}_{2,p} < c_\nu$ and, thus, $P^{(i)}(\mathbf{Y} \in t_L) = \Phi\left(\frac{\grave{c}_{2,p} - \mu_{Y,\nu,i}}{\sqrt{\sigma_{\nu\nu}}}\right)$, and $P^{(i)}(\mathbf{Y} \in t_R) = \Phi\left(\frac{c_\nu - \mu_{Y,\nu,i}}{\sqrt{\sigma_{\nu\nu}}}\right) - \Phi\left(\frac{\grave{c}_{2,p} - \mu_{Y,\nu,i}}{\sqrt{\sigma_{\nu\nu}}}\right)$. We then let c_κ denote the cutpoint $\grave{c}_{2,p}$ for which (4.9) or (4.10) are minimized or (4.8) is maximized across all p and let Y_κ denote the feature variable corresponding to this particular cutpoint. The right daughter node t_R of t_0 is split in the same manner.

We continue to split all subsequent descendant nodes in this fashion until some stopping criterion is met, as described in Section 4.1.2.1, and use the rule in (4.2) to assign each terminal node in our tree T' to one of the two groups. Using the MATLAB® software package, we have been able to construct T' in this manner based on a number of examples, one of which deals with the case of $p = 6$.

4.1.2.4 Monotone Invariance Property We now discuss the following important known property of the method used in the construction of classification trees, which we carefully prove for our purposes.

Proposition 4.1.2.1. *Based on either the impurity measure based GOS criterion or the twoing criterion, let T' be the classification tree based on the priors $\pi_1, \ldots, \pi_g$ and the distribution functions $F_{\mathbf{Y}}^{(1)}(\cdot)$, $\ldots$, $F_{\mathbf{Y}}^{(g)}(\cdot)$ $(g \geq 2)$. Further, let $\mathbf{Z} = (Z_1, \ldots, Z_P)' = (\zeta_1(Y_1), \ldots, \zeta_P(Y_P))' \equiv \boldsymbol{\zeta}(\mathbf{Y})$, where $\zeta_p(Y_p)$ is a strictly increasing function of Y_p $(p = 1, \ldots, P)$ or, in other words, $\boldsymbol{\zeta}(\mathbf{Y})$ is a monotonic transformation of $\mathbf{Y}$. Let $T'_{\mathbf{Z}}$ be the classification tree based on the priors $\pi_1, \ldots, \pi_g$ and the distribution functions $G_{\mathbf{Z}}^{(1)}(\cdot)$, $\ldots$, $G_{\mathbf{Z}}^{(g)}(\cdot)$. Then, T' and $T'_{\mathbf{Z}}$ have the same structure, with the set of splitting variables for T', $\mathbf{Y}_{T'}$, related to those of $T'_{\mathbf{Z}}$, $\mathbf{Z}_{T'_{\mathbf{Z}}}$, by $Z_{T'_{\mathbf{Z}},p} = \zeta_p(Y_{T',p})$ $(p = 1, \ldots, P)$ and the set of cutpoints for T', $\mathbf{c}_{T'}$, related to those of $T'_{\mathbf{Z}}$, $\mathbf{c}_{T'_{\mathbf{Z}}}$, by $\mathbf{c}_{T'_{\mathbf{Z}}} = \boldsymbol{\zeta}(\mathbf{c}_{T'})$.*

Proof. The proof proceeds by induction.

Suppose that Y_ν and c_ν are chosen to split the root node t_0 into descendant nodes t_L and t_R, such that the selected GOS criterion defined by the split $Y_\nu \leq c_\nu$ is maximized. In our discussion of the GOS criteria in Section 4.1.2.2, we noted that whether we seek to minimize the formula in (4.6) based on the impurity measure $M(t)$ or maximize the twoing criterion over each Y_p $(p = 1, \ldots, P)$ in $\mathbf{Y}$ and all cutpoints $c \in \mathbb{R}$, both quantities can be expressed as functions of π_i $(i = 1, \ldots, g)$, $P(\mathbf{Y} \in t_0) = 1$,

$$\begin{aligned} P^{(i)}(\mathbf{Y} \in t_L) &= P^{(i)}(Y_\nu \leq c_\nu) \\ &= P^{(i)}(\zeta_\nu(Y_\nu) \leq \zeta_\nu(c_\nu)) \\ &= P^{(i)}(Z_\nu \leq \zeta_\nu(c_\nu)), \end{aligned}$$

and

$$\begin{aligned} P^{(i)}(\mathbf{Y} \in t_R) &= P^{(i)}(Y_\nu > c_\nu) \\ &= P^{(i)}(\zeta_\nu(Y_\nu) > \zeta_\nu(c_\nu)) \\ &= P^{(i)}(Z_\nu > \zeta_\nu(c_\nu)). \end{aligned}$$

For example, based on the Gini index $M_G(t)$, we have that the formula in (4.6) is equal to

$$\begin{aligned} &2\pi_1\pi_2 \left[\frac{F^{(1)}_{Y_\nu}(c_\nu) F^{(2)}_{Y_\nu}(c_\nu)}{\pi_1 F^{(1)}_{Y_\nu}(c_\nu) + \pi_2 F^{(2)}_{Y_\nu}(c_\nu)} + \frac{\bar{F}^{(1)}_{Y_\nu}(c_\nu) \bar{F}^{(2)}_{Y_\nu}(c_\nu)}{\pi_1 \bar{F}^{(1)}_{Y_\nu}(c_\nu) + \pi_2 \bar{F}^{(2)}_{Y_\nu}(c_\nu)} \right] \\ &= 2\pi_1\pi_2 \left[\frac{G^{(1)}_{Z_\nu}(\zeta_\nu(c_\nu))\, G^{(2)}_{Z_\nu}(\zeta_\nu(c_\nu))}{\pi_1 G^{(1)}_{Z_\nu}(\zeta_\nu(c_\nu)) + \pi_2 G^{(2)}_{Z_\nu}(\zeta_\nu(c_\nu))} + \frac{\bar{G}^{(1)}_{Z_\nu}(\zeta_\nu(c_\nu))\, \bar{G}^{(2)}_{Z_\nu}(\zeta_\nu(c_\nu))}{\pi_1 \bar{G}^{(1)}_{Z_\nu}(\zeta_\nu(c_\nu)) + \pi_2 \bar{G}^{(2)}_{Z_\nu}(\zeta_\nu(c_\nu))} \right], \end{aligned}$$

for two groups, where $F^{(i)}_Y(c) = P^{(i)}(Y \leq c)$, $G^{(i)}_Z(\zeta(c)) = P^{(i)}(Z \leq \zeta(c))$, $\bar{F}^{(i)}_Y(c) = 1 - F^{(i)}_Y(c)$, and $\bar{G}^{(i)}_Y(\zeta(c)) = 1 - G^{(i)}_Y(\zeta(c))$. Therefore, if Y_ν and c_ν are first chosen to split the feature space $\mathcal{Y}$ in the construction of tree T', then Z_ν and $\zeta_\nu(c_\nu)$ are first chosen to split $\mathcal{Z}$, the monotonic transformation of $\mathcal{Y}$, in the construction of tree $T'_{\mathbf{Z}}$.

Suppose now that we are in a step of the algorithm where there are m descendant nodes or subsets of $\{\mathcal{Y} : Y_\nu \leq c_\nu\}$ in T' and $\{\mathcal{Z} : Z_\nu \leq \zeta_\nu(c_\nu)\}$ in $T'_{\mathbf{Z}}$, as well as the m' descendant subsets of $\{\mathcal{Y} : Y_\nu > c_\nu\}$ in T' and $\{\mathcal{Z} : Z_\nu > \zeta_\nu(c_\nu)\}$ in $T'_{\mathbf{Z}}$. By the induction setup, we assume that if the split $Y_\upsilon \leq c_\upsilon$ is used for a particular node t in T', then the split $Z_\upsilon \leq \zeta_\upsilon(c_\upsilon)$ is used for the corresponding node $t_{\mathbf{Z}}$ in $T'_{\mathbf{Z}}$.

Let Y_κ and c_κ be chosen to split the $(m+1)^{st}$ descendant node t of $\{\mathcal{Y} : Y_\nu \leq c_\nu\}$, into daughter nodes t_L and t_R, such that the selected GOS criterion defined by the split $Y_\kappa \leq c_\kappa$ is maximized. Using the same procedure as that used to split t_0, we can conclude that if Y_κ and c_κ are chosen to split the $(m+1)^{st}$ descendant node of $\{\mathcal{Y} : Y_\nu \leq c_\nu\}$ in the construction of tree T', then Z_κ and $\zeta_\kappa(c_\kappa)$ are chosen to split the $(m+1)^{st}$ descendant node of $\{\mathcal{Z} : Z_\nu \leq \zeta_\nu(c_\nu)\}$ in the construction of tree $T'_{\mathbf{Z}}$. The same result holds if we wish to split the $(m'+1)^{st}$ descendant node of $\{\mathcal{Y} : Y_\nu > c_\nu\}$.

Thus, by induction, we have that $\mathbf{Z}_{T'_{\mathbf{Z}}} = \boldsymbol{\zeta}(\mathbf{Y}_{T'})$ and $\mathbf{c}_{T'_{\mathbf{Z}}} = \boldsymbol{\zeta}(\mathbf{c}_{T'})$. □

Based on Proposition 4.1.2.1, classification trees are invariant under all monotonic coordinate-wise transformations of $\mathbf{Y}$. For example, suppose we consider the trees T' and $T'_{\mathbf{Z}}$ based on $\mathbf{Y}$ and the monotonic transformation $\mathbf{Z} \equiv \boldsymbol{\zeta}(\mathbf{Y})$, respectively. We have that the split $Y_\nu \leq c_\nu$ $(\nu \in (1, 2, \ldots, P))$ in T' is equivalent to the split $\zeta_\nu(Y_\nu) \leq \zeta_\nu(c_\nu)$ in $T'_{\mathbf{Z}}$. Specifically, the observations that fall in the left descendant node of the split $Y_\nu \leq c_\nu$ are the same as those that fall in the left descendant node of the split $\zeta_\nu(Y_\nu) \leq \zeta_\nu(c_\nu)$ and likewise for the right descendant nodes of these splits. Therefore, T' and $T'_{\mathbf{Z}}$ have the same classification results and if T' identifies a set of discriminatory feature variables, then $T'_{\mathbf{Z}}$ identifies the same set, transformed by the function $\boldsymbol{\zeta}(\centerdot)$.

4.1.3 Estimation of Unknown Distributions using Training Data

Suppose that we only have access to training data consisting of the observed feature measurements $\mathbf{y}_{ij}$ $(i = 1, \ldots, g; j = 1, \ldots, n_i)$, where $N = \sum_{i=1}^{g} n_i$ is the total sample size. We note that this is the traditional setting in which the BFOS recursive partitioning algorithm was developed. The prior probabilities π_i are typically specified in advance or sometimes estimated from the training data.

There are a few approaches we can take to carry out the tree construction procedure in Section 4.1.2.1 using the training data. If one were to assume known distributions with unknown parameters, then we suggest applying our population-based extension of the BFOS algorithm to this parametric case, after estimating the parameters from the training data. On the other hand, if no distributional assumptions are made, $F^{(i)}_{\mathbf{Y}}(\centerdot)$ can be estimated non-parametrically using either empirical CDFs, as is done in the traditional BFOS algorithm in

the data setting, or kernel density estimation, which Shang and Breiman consider in their tree construction methodology [31].

4.1.3.1 Parametric Approach We can assume that, for the i^{th} group, $\mathbf{Y}$ comes from some parametric distribution and use the training data to estimate any unknown parameters. Once we obtain the estimated parameters, we can use the assumed distribution of $\mathbf{Y}$ to obtain the estimated probabilities $\hat{P}^{(i)}(\mathbf{Y} \in t)$, $\hat{P}^{(i)}(\mathbf{Y} \in t_L)$, and $\hat{P}^{(i)}(\mathbf{Y} \in t_R)$. The tree construction procedure can then proceed as in Section 4.1.2.1, terminating only when the number of observations in node t is less than some user defined value.

For example, we might assume that $\mathbf{Y} \sim N_P(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\Sigma}_{YY})$ in the i^{th} group, where $\boldsymbol{\mu}_{Y,i}$ and $\boldsymbol{\Sigma}_{YY}$ are unknown. We can then use the training data to obtain the ML estimates $\hat{\boldsymbol{\mu}}_{Y,i} = \bar{\mathbf{y}}_i$ and $\hat{\boldsymbol{\Sigma}}_{YY}$ and construct our tree using the procedure described in Section 4.1.2.3, which, as we indicated, we have been able to implement in MATLAB®.

4.1.3.2 Non-parametric Approach The standard approach used in constructing classification trees is to non-parametrically estimate the CDFs $F^{(i)}_{\mathbf{Y}}(\cdot)$ $(i = 1, \ldots, g)$ and use them in the tree construction procedure described in Section 4.1.2.1 [7][13][14][30]. Although Shang and Breiman proposed a method to estimate $F^{(i)}_{\mathbf{Y}}(\cdot)$ using kernel density estimation [31], the usual approach is to compute the empirical CDF, $\hat{F}^{(i)}_{\mathbf{Y}}(\cdot)$, of $\mathbf{Y}$ in the i^{th} group, where

$$\hat{F}^{(i)}_{\mathbf{Y}}(\mathbf{c}) = \hat{P}^{(i)}(Y_1 \leq c_1, \ldots, Y_P \leq c_P) = \frac{\sum_{j=1}^{n_i} I(y_{ij,1} \leq c_1, \ldots, y_{ij,P} \leq c_P)}{n_i} \tag{4.11}$$

and $I(\cdot)$ is the indicator function. Although we assume in our discussion that $\mathbf{Y}$ is continuous, the estimate in (4.11) is applicable for any quantitative feature vector.

In general, $P^{(i)}(\mathbf{Y} \in t)$ can then be estimated as

$$\hat{P}^{(i)}(\mathbf{Y} \in t) = \frac{\sum_{j=1}^{n_i} I(\mathbf{y}_{ij} \in t)}{n_i}, \tag{4.12}$$

the sample proportion of feature observations in group i that fall in node t. BFOS prove that, under appropriate conditions, the group assignment rule in (4.2) based on π_i and $\hat{P}^{(i)}(\mathbf{Y} \in t)$ is Bayes risk consistent [7]. In other words, as the sample sizes n_i approach

infinity, the estimated misclassification rate for node t based on the training data converges in probability to the Bayes or true misclassification rate for node t in (4.3).

Once we obtain π_i and the probability estimates in (4.12), the tree construction procedure can then proceed as in Section 4.1.2.1. In particular, splitting continues until all feature measurements in t belong to the same group or are identical, or the number of observations in t is less than some user defined value [7][23]. It is important to note that this non-parametric procedure can be readily implemented using various software packages, e.g., R or Salford Systems CART®. In addition, the procedure described in this section has been extended to handle missing data [7][14].

4.1.3.3 Misclassification Rate Estimates Let T_{max} denote a tree obtained using either the parametric or non-parametric approach and $\tilde{T}_{max}$ denote the set of terminal nodes of T_{max}. Since the distribution of $\mathbf{Y}$ is unknown for each group, we can no longer determine $R^*(T_{max})$, the true misclassification rate for T_{max}. One estimate of $R^*(T_{max})$ is the resubstituted or plug in estimate

$$\hat{R}^*(T_{max}) = \sum_{t \in \tilde{T}_{max}} \left\{ \hat{P}(\mathbf{Y} \in t) - max_{i=1,\ldots,g} \left[\pi_i \hat{P}^{(i)}(\mathbf{Y} \in t) \right] \right\},$$

where $\hat{P}(\mathbf{Y} \in t) = \sum_{i=1}^{g} \pi_i \hat{P}^{(i)}(\mathbf{Y} \in t)$. However, one problem with using $\hat{R}^*(T)$ to estimate the true misclassification rate $R^*(T)$ for tree T is that it is computed using the same sample that was used to construct T, instead of an independent sample. Thus, $\hat{R}^*(T_{max})$ is likely to be overly optimistic in estimating the accuracy of T_{max}. Furthermore, it is known that $\hat{R}^*(T)$ becomes increasingly less accurate as the T grows larger in size, where the size (or complexity) of T is the number of terminal nodes in T [7][23]. On the other hand, estimates of the true misclassification rate for T that are obtained from independent sampling techniques, i.e., test sampling or V-fold cross validation ($V = 2, \ldots, N$), have been shown to be more accurate and less biased compared to those obtained from resubstitution [7][13][23].

4.1.3.4 Minimal Cost-Complexity Pruning Several authors highlight the fact that if T_{max} is constructed using the standard non-parametric approach, then T_{max} substantially overfits the training data, which contributes to splits at the lower levels of T_{max} being determined mainly by sampling fluctuations rather than actual underlying data structures for the

groups of interest [7][13][23]. Thus, T_{max} cannot be generalized to new data in this case. In their attempt to solve this problem, BFOS devised a backward node recombination strategy specific to their non-parametric tree construction procedure called minimal cost complexity pruning [7]. Prior to presenting this method, we introduce some notation. A branch of tree T consists of a parent node t in T and all descendant nodes of t. Pruning a branch with parent node t from T entails cutting off all descendant nodes of t. A subtree of T is then obtained when a particular branch or branches are pruned from T. The cost-complexity of T is defined as $C_\alpha(T) = \hat{R}^*(T) + \alpha \mid \tilde{T} \mid$, where $\hat{R}^*(T)$ is the observed or resubstituted misclassification rate for tree T, $\mid \tilde{T} \mid$ is the size of T, and α is a positive real number called the complexity parameter.

The goal of minimal cost-complexity pruning is to prune from T_{max} the weakest-link branch or branches necessary to obtain a subtree T_* of T_{max} that has the smallest cost complexity. BFOS prove that for every α value, there exists a unique smallest subtree $T(\alpha)$ of T_{max} that minimizes $C_\alpha(T)$ [7]. For example, if $\alpha = 0$, then $T(0) = T_1$, where T_1 is the smallest subtree of T_{max} such that $\hat{R}^*(T_1) = \hat{R}^*(T_{max})$. Increasing α yields the nested sequence of optimal subtrees of T_{max} of decreasing size $T_1 \supset T_2 \supset \cdots \supset \{t_0\}$, each of which is the best tree of its size, i.e., has the smallest cost complexity [7][13][23]. T_* is the best subtree in this sequence and is chosen on the basis of misclassification rate estimates computed using either test sampling or V-fold cross validation techniques, both of which can be carried out using any software package that can implement the standard non-parametric approach described in Section 4.1.3.2.

We begin with a brief description of test sampling, which starts out by randomly selecting $N^{(ts)}$ of the N individuals in the training data to constitute the test sample and the remaining $N - N^{(ts)}$ individuals the learning sample. T_{max} is constructed using only the learning sample data and is then pruned upward to yield the nested sequence of optimal subtrees $T_1 \supset T_2 \supset \cdots \supset \{t_0\}$. Each of the trees in this sequence is used to classify all of the individuals in the test sample. For each subtree T_m $(m = 1, 2, \dots)$, we let $R_m^{(ts)}(j)$ denote the proportion of group j $(j = 1, \dots, g)$ test cases that are misclassified, after which the test sample misclassification rate estimate for T_m is computed as $R^{(ts)}(T_m) = \sum_{j=1}^{g} \pi_j R_m^{(ts)}(j)$. The subtree T_* of T_{max} is chosen such that $R^{(ts)}(T_*) = min_m\, R^{(ts)}(T_m)$.

On the other hand, V-fold cross validation begins by constructing T_{max} using the *entire*

training data set. T_{max} is then pruned upward to yield the nested sequence of optimal subtrees $T_1 \supset T_2 \supset \cdots \supset \{t_0\}$, from which we must find the subtree T_* with the smallest cost complexity. Once T_{max} is pruned in this manner, the training data is randomly split into V ($V = 2, \ldots, N$) mutually exclusive subsets of approximately equal size, stratified by group. Each of the V subsets is dropped out, while the tree T^v_{max} ($v = 1, \ldots, V$) is computed using the remaining $V - 1$ subsets, so that V additional trees are constructed along with T_{max}. Using the same procedure as that used for T_{max}, T^v_{max} is pruned upward to yield the nested sequence of optimal trees of decreasing size $T^v_1 \supset T^v_2 \cdots \supset \{t_0\}$. We note here that the number of subtrees in this sequence is the same as that for T_{max}. Each of the subtrees $T^v_1, T^v_2, \ldots$ are then used to classify each individual in the omitted subset. For each subtree T^v_m, the number of group j individuals in the v^{th} subset that are misclassified as belonging to group i ($i, j = 1, \ldots, g;\ i \neq j$) is computed and added across all V subsets, with this sum denoted as N^m_{ij}. With regards to the nested sequence $T_1 \supset T_2 \supset \cdots \supset \{t_0\}$ for T_{max}, the cross validated misclassification rate estimate for subtree T_m is computed as $R^{(cv)}(T_m) = \sum_{j=1}^{g} \sum_{\substack{i=1 \\ i \neq j}}^{g} \pi_j \frac{N^m_{ij}}{n_j}$. The subtree T_* of T_{max} is then chosen such that $R^{(cv)}(T_*) = min_m\, R^{(cv)}(T_m)$.

4.2 CONDITIONAL CLASSIFICATION TREES

4.2.1 Motivation

From our discussion of traditional classification trees in the previous section, we saw that the tree construction procedure used to partition the feature space $\mathcal{Y}$ is solely based on the feature data. However, this procedure does not account for the relationship that may exist between the feature data and other relevant covariates. For example, in the context of post-mortem tissue studies that compare schizophrenia subjects with normal controls, traditional classification trees would not account for the relationship that typically exists between a particular set of biomarkers and covariates such as storage time or brain pH. Thus, we cannot be confident that such trees will help us determine the subset of biomarkers, as well as the corresponding splits on these biomarkers, that truly discriminate between the control and

schizophrenia diagnostic groups. In our motivation of covariance adjusted LDA, we showed how examining the conditional distribution of the feature data while holding covariates fixed allows us to accurately determine the true discriminatory power of the feature data. We now develop the approach to extend the ideas of the BFOS recursive partitioning algorithm in order to adjust for covariate effects in the construction of classification trees.

In their development of covariance adjusted LDA, Cochran and Bliss [8], Lachenbruch [19], and Tu et al. [37] assume that, if all covariates are held fixed, the feature data come from a normal population for each group with a known conditional mean and variance-covariance matrix that is common across groups.

Our goal is to extend the traditional BFOS classification tree construction procedure to account for the effects of covariates on the feature data, which would help us achieve our primary goal of determining the subset of feature variables that best discriminate between the groups of interest without the confounding effects of any covariates under consideration. We again follow the approach used in LDA in that we begin from a population based standpoint and then apply our results to the case where we only have access to training data. Our development of conditional classification trees first considers g known conditional continuous distributions for the feature data. Although we show in Section 4.2.2.2 that our methodology for constructing conditional classification trees can be developed for normal populations, our focus is on the case of g known arbitrary conditional distributions. We then discuss how our methodology can be implemented using available training data in Section 4.2.3.

4.2.2 Known Conditional Distributions

4.2.2.1 Tree Construction Procedure Given $\mathbf{X} = \mathbf{x}$, let $\mathbf{Y}$ have some known conditional CDF $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ in the i^{th} group ($i = 1, \ldots, g$). For a given $\mathbf{x}$, $\mathbf{Y}$ conceptually could also have a known prior probability $\pi_i(\mathbf{x})$ of belonging to group i, depending on the covariate value $\mathbf{x}$. In other words, given $\mathbf{X} = \mathbf{x}$, the conditional distribution of $\mathbf{Y}$ is a mixture of the conditional distributions of $\mathbf{Y}$ across the g groups, so that

$$F_{\mathbf{Y}|\mathbf{x}}(\centerdot) = \sum_{i=1}^{g} \pi_i(\mathbf{x}) F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot). \tag{4.13}$$

We also retain our assumption from Section 4.1.2.1 of equal misclassification costs.

In our construction of the tree T' in Section 4.1.2.1, the GOS criteria and group assignment rule were primarily expressed as functions of probabilities of the form $P(\mathbf{Y} \in t)$ and $P(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t)$. However, we were able to re-express these probabilities in terms of π_i and $P^{(i)}(\mathbf{Y} \in t)$, which are computed using the CDF of $\mathbf{Y}$ in the i^{th} group $F^{(i)}_{\mathbf{Y}}(\centerdot)$ $(i = 1, \ldots, g)$, i.e., the distribution of $\mathbf{Y}$ is a mixture of the marginal distributions of $\mathbf{Y}$ across all g groups.

For a given $\mathbf{x}$, we replace $P(\mathbf{Y} \in t)$ and $P(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t)$ with their conditional counterparts $P_{\mathbf{x}}(\mathbf{Y} \in t) = \frac{1}{f(\mathbf{x})} \int_{\mathbf{y} \in t} f(\mathbf{x}, \mathbf{y}) d\mathbf{y}$ and $P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t) = \frac{\int_{\mathbf{y} \in t, \mathbf{y} \in \text{group } i} f(\mathbf{x},\mathbf{y}) d\mathbf{y}}{\int_{\mathbf{y} \in t} f(\mathbf{x},\mathbf{y}) d\mathbf{y}}$ in our construction of the conditional tree $T'^{(\mathbf{x})}$. For notational convenience, we assume in the above formulas and throughout our discussion that all relevant random variables are continuous with existing density functions.

We now describe how to construct $T'^{(\mathbf{x})}$ based on our knowledge of the conditional CDFs $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ and priors $\pi_i(\mathbf{x})$ for a given $\mathbf{x}$. First, we point out that the conditional probabilities $P_{\mathbf{x}}(\mathbf{Y} \in t)$ and $P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t)$ can also be expressed as

$$\begin{aligned} P_{\mathbf{x}}(\mathbf{Y} \in t) &= \sum_{i=1}^{g} P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i, \mathbf{Y} \in t) \\ &= \sum_{i=1}^{g} P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i) P_{\mathbf{x}}(\mathbf{Y} \in t | \mathbf{Y} \in \text{group } i) = \sum_{i=1}^{g} \pi_i(\mathbf{x}) P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t) \end{aligned} \tag{4.14}$$

and

$$P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t) = \frac{P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i, \mathbf{Y} \in t)}{P_{\mathbf{x}}(\mathbf{Y} \in t)} = \frac{\pi_i(\mathbf{x}) P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t)}{\sum_{j=1}^{g} \pi_j(\mathbf{x}) P^{(j)}_{\mathbf{x}}(\mathbf{Y} \in t)}, \tag{4.15}$$

where $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t) = P_{\mathbf{x}}(\mathbf{Y} \in t | \mathbf{Y} \in \text{group } i)$ are computed using the conditional CDFs $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ $(i = 1, \ldots, g)$. We see that the formulas in (4.14) and (4.15) are identical to those in (4.4) and (4.1), respectively, for $P(\mathbf{Y} \in t)$ and $P(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t)$, where π_i and $P^{(i)}(\mathbf{Y} \in t)$ are replaced with $\pi_i(\mathbf{x})$ and $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t)$. Therefore, assuming that the conditional CDFs $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ and prior probabilities $\pi_i(\mathbf{x})$ are known for a given $\mathbf{x}$, we can construct our tree $T'^{(\mathbf{x})}$ in the same manner as that used to construct the traditional classification tree T' in Section 4.1.2.1 by simply replacing the probabilities π_i, $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ used to construct T' with their conditional counterparts $\pi_i(\mathbf{x})$, $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t)$, $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t_L)$, and $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t_R)$.

Alternately, we can have that the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ and the marginal distribution of $\mathbf{X}$ are known, rather than the conditional distribution of $\mathbf{Y}$. To elaborate, we let the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ be a mixture of the joint distributions of $\mathbf{X}$ and $\mathbf{Y}$ across all g groups so that

$$f(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{g} \pi_i^*(\mathbf{x}) f_i(\mathbf{x},\mathbf{y}), \tag{4.16}$$

where the prior probabilities $\pi_i^*(\mathbf{x})$ and the densities $f_i(\mathbf{x},\mathbf{y})$ are known, and, thus,

$$f(\mathbf{x}) = \sum_{i=1}^{g} \pi_i^*(\mathbf{x}) \int_{-\infty}^{\infty} f_i(\mathbf{x},\mathbf{y}) d\mathbf{y} = \sum_{i=1}^{g} \pi_i^*(\mathbf{x}) f_i(\mathbf{x}).$$

When constructing our conditional tree $T^{'(\mathbf{x})}$ in practice, we want the prior probability that $\mathbf{Y}$ given $\mathbf{x}$ belongs to the i^{th} group to be a function of $\mathbf{x}$ that does not change depending on whether we use the model in (4.13) for the conditional distribution of $\mathbf{Y}$ or (4.16) for the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$. The reason we want this to hold at this stage in our development of conditional classification trees is that we do not want to be concerned with the specifics of the data sampling methods. Therefore, we discuss certain conditions that must hold so that the prior probabilities $\pi_i^*(\mathbf{x})$ in (4.16) are the same functions of $\mathbf{x}$ as the prior probabilities $\pi_i(\mathbf{x})$ in (4.13), i.e., $\pi_i^*(\mathbf{x}) = \pi_i(\mathbf{x})$. Specifically, based on the model in (4.16),

$$f(\mathbf{y}|\mathbf{x}) = \frac{1}{f(\mathbf{x})} \sum_{i=1}^{g} \pi_i^*(\mathbf{x}) f_i(\mathbf{x},\mathbf{y}) = \frac{1}{f(\mathbf{x})} \sum_{i=1}^{g} \pi_i^*(\mathbf{x}) f_i(\mathbf{y}|\mathbf{x}) f_i(\mathbf{x}) = \sum_{i=1}^{g} \pi_i^{**}(\mathbf{x}) f_i(\mathbf{y}|\mathbf{x}), \tag{4.17}$$

where $\pi_i^{**}(\mathbf{x}) = \frac{\pi_i^*(\mathbf{x}) f_i(\mathbf{x})}{f(\mathbf{x})}$. We have that the probabilities $\pi_i^{**}(\mathbf{x})$ associated with $f_i(\mathbf{y}|\mathbf{x})$ in (4.17) are mixture weights since $\pi_i^{**}(\mathbf{x}) \geq 0$ and

$$\sum_{i=1}^{g} \pi_i^{**}(\mathbf{x}) = \frac{\sum_{i=1}^{g} \pi_i^*(\mathbf{x}) f_i(\mathbf{x})}{f(\mathbf{x})} = \frac{f(\mathbf{x})}{f(\mathbf{x})} = 1.$$

From (4.17), $F_{\mathbf{Y}|\mathbf{x}}(\centerdot) = \sum_{i=1}^{g} \pi_i^{**}(\mathbf{x}) F_{\mathbf{Y}|\mathbf{x}}^{(i)}(\centerdot)$, which is equivalent to the model in (4.13) if and only if $\pi_i^{**}(\mathbf{x}) = \pi_i(\mathbf{x})$. In addition, based on the formula for $\pi_i^{**}(\mathbf{x})$, we have that $\pi_i^{**}(\mathbf{x}) = \pi_i^*(\mathbf{x})$ if $f_i(\mathbf{x}) \equiv f(\mathbf{x})$. Thus, if $\pi_i^{**}(\mathbf{x}) = \pi_i(\mathbf{x})$ and $f_i(\mathbf{x}) \equiv f(\mathbf{x})$, then $\pi_i^*(\mathbf{x}) = \pi_i^{**}(\mathbf{x}) = \pi_i(\mathbf{x})$ and the model in (4.13) can equivalently be obtained from the model in (4.16). In particular, if $\pi_i^{**}(\mathbf{x}) = \pi_i(\mathbf{x})$ and $f_i(\mathbf{x}) \equiv f(\mathbf{x})$, then the prior probabilities and conditional probabilities

$P_{\mathbf{x}}(\mathbf{Y} \in t)$ and $P_{\mathbf{x}}(\mathbf{Y} \in \text{group } i | \mathbf{Y} \in t)$ needed to construct $T'^{(\mathbf{x})}$ for a given $\mathbf{x}$ remain the same under each of the models in (4.13) and (4.16).

It is important to note that the conditional tree $T'^{(\mathbf{x})}$ varies depending on $\mathbf{x}$. In other words, the feature variable Y in $\mathbf{Y}$ and the cutpoint c chosen to split any node t in $T'^{(\mathbf{x})}$ both depend on the value at which $\mathbf{x}$ is fixed. Although it is not problematic in the context of classification that the cutpoints for $T'^{(\mathbf{x})}$ are covariate dependent, the fact that the set of splitting variables chosen for $T'^{(\mathbf{x})}$ changes depending on the value of $\mathbf{x}$ is not a desirable property in certain contexts, such as those pertaining to post-mortem tissue studies. Therefore, we want to develop a model for the conditional distribution of $\mathbf{Y}$ for a given $\mathbf{x}$ or the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$, depending on which is known, such that the set of feature variables selected in the construction of $T'^{(\mathbf{x})}$ remains the same regardless of the value of $\mathbf{x}$, although cutpoints will vary.

In the spirit of Lachenbruch and Tu et al. and for ease of computation, we only consider the conditional model in (4.13) from this point on in our construction of $T'^{(\mathbf{x})}$. Based on the monotone invariance property (Proposition 4.1.2.1), we have that the set of feature variables chosen for $T'^{(\mathbf{x})}$ does not depend on $\mathbf{x}$ if $\pi_i(\mathbf{x}) \equiv \pi_i$ regardless of $\mathbf{x}$ and the conditional distribution of $\mathbf{Y}$ belongs to a location-scale family, which we show in greater detail in Section 4.3.1.1. The assumption that $\pi_i(\mathbf{x}) \equiv \pi_i$ is of practical relevance in our development of conditional classification trees. Specifically, having the prior probabilities be covariate dependent implies that $\mathbf{X}$ also has discriminatory importance, which we assume in our discussion is not the case. Rather, we only want to account or control for covariate effects in order to more accurately identify the feature variables in $\mathbf{Y}$ with the highest discriminatory importance. For example, even though tissue storage time or brain pH may differ somewhat between schizophrenia subjects and normal controls in post-mortem tissue studies, we have no interest in including these covariates in our discrimination, other than to control for their effects on the biomarker data.

In general, the methodology we formulate in this section allows us to construct a tree that adjusts for the effects of the covariate vector $\mathbf{X}$, while still using the traditional tree construction approach described in Section 4.1.2.1.

4.2.2.2 Tree Construction for Known Normal Populations

Given $\mathbf{X} = \mathbf{x}$, suppose $\mathbf{Y} = (Y_1, \ldots, Y_P)'$ has prior probability $\pi_i(\mathbf{x}) \equiv \pi_i$ of belonging to the i^{th} group $(i = 1, \ldots, g)$, where $\mathbf{Y} \sim N_P(\boldsymbol{\varsigma}_i(\mathbf{x}), \boldsymbol{\Sigma}_{\mathbf{x}})$ and $\boldsymbol{\varsigma}_i(\mathbf{x}) = (\varsigma_{i,1}(\mathbf{x}), \ldots, \varsigma_{i,P}(\mathbf{x}))'$ is some known function of $\mathbf{x}$. We note that this is the conditional model on which general covariance adjusted LDA is based. First, we split the root node t_0 into t_L and t_R and use the fact that given $\mathbf{X} = \mathbf{x}$, $\frac{Y_p - \varsigma_{i,p}(\mathbf{x})}{\sqrt{\sigma_{pp(\mathbf{x})}}} \sim N(0,1)$ for the i^{th} group, where $\sigma_{pp(\mathbf{x})}$ denotes the conditional variance of Y_p $(p = 1, \ldots, P)$. We then proceed to find, for each feature variable Y_p, the cutpoint $\grave{c}_{1,p}$ that minimizes the quantities in (4.9) or (4.10) or maximizes the quantity in (4.8), depending on the GOS criterion and impurity measure that is chosen. For these three formulas, we replace $P(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ with the conditional probabilities $P_{\mathbf{x}}(\mathbf{Y} \in t)$, $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t_L)$, and $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t_R)$. In the first step of our tree construction procedure, $P_{\mathbf{x}}(\mathbf{Y} \in t) = P_{\mathbf{x}}(\mathbf{Y} \in t_0) = 1$, since t_0 is the entire feature space. For each Y_p and $\grave{c}_{1,p}$, $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t_L) = P^{(i)}_{\mathbf{x}}(Y_p \leq \grave{c}_{1,p}) = \Phi\left(\frac{\grave{c}_{1,p} - \varsigma_{i,p}(\mathbf{x})}{\sqrt{\sigma_{pp(\mathbf{x})}}}\right)$ and $P^{(i)}_{\mathbf{x}}(\mathbf{Y} \in t_R) = P^{(i)}_{\mathbf{x}}(Y_p > \grave{c}_{1,p}) = \Phi\left(\frac{\varsigma_{i,p}(\mathbf{x}) - \grave{c}_{1,p}}{\sqrt{\sigma_{pp(\mathbf{x})}}}\right)$. We then let c_ν denote the cutpoint $\grave{c}_{1,p}$ that minimizes (4.9) or (4.10) or maximizes (4.8) across all p and let Y_ν denote the feature variable that corresponds to this particular cutpoint, noting that Y_ν and c_ν now depend on $\mathbf{x}$.

Suppose we consider the case of two groups, such that $\mathbf{Y} = (Y_1, \ldots, Y_P)' \sim N_P(\boldsymbol{\mu}_{Y|X,i}, \boldsymbol{\Sigma}_{Y|X})$ in the i^{th} group $(i = 1, 2)$, where $\pi_i = 0.5$ and $\boldsymbol{\mu}_{Y|X,i}$ and $\boldsymbol{\Sigma}_{Y|X}$ are defined as in Section 3.2.2 for traditional covariance adjusted LDA. We may also examine the covariate adjusted feature vector $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \ldots, \tilde{Y}_P)' \sim N_P(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\Sigma}_{Y|X})$ in the i^{th} group, where $\tilde{\mathbf{Y}}$ is defined as in Section 3.2.2. If we use the differentiability of the Gini and Deviance indices and the twoing criterion, then it can be shown that the optimal cutpoint $\tilde{\grave{c}}_{1,p}$ for each $\tilde{Y}_p$ is equal to $\frac{1}{2}(\mu_{Y,p,1} + \mu_{Y,p,2})$. In particular, if $\tilde{\mathbf{Y}}$ is univariate such that $\tilde{\mathbf{Y}} = \tilde{Y}_1$, the optimal cutpoint $\frac{1}{2}(\mu_{Y,1,1} + \mu_{Y,1,2})$ corresponding to $\tilde{Y}_1$ in the first split of our covariate adjusted tree is the same cutpoint used to split the feature space $\mathcal{Y}$ in traditional covariance adjusted LDA for two groups. If we were to stop splitting at this point, we would have that the true misclassification rates for our resulting adjusted tree and the classification rule given in (3.8) in Section 3.2.2 would be identical. Under the assumptions of this two group case, it can also be shown that if the feature space $\mathcal{Y}$ is only split once, then our covariate adjusted tree would yield a lower misclassification rate than the traditional tree T', assuming the distribution of $\mathbf{Y}$ is also normal in each of the two groups. When we deal with more than two groups, however, the

results obtained from traditional covariance adjusted LDA and the BFOS algorithm based on univariate $\tilde{\mathbf{Y}}$ when we first split the feature space can be shown to no longer be the same.

In the general case of g groups, the splitting procedure described in the first step is then applied recursively to t_L, t_R, and all subsequent descendant nodes until some stopping criterion is satisfied, as in Section 4.1.2.3 in the traditional case. The rule in (4.2), where $P^{(i)}(\mathbf{Y} \in t)$ are replaced with $P_{\mathbf{x}}^{(i)}(\mathbf{Y} \in t)$ $(i = 1, \ldots, g)$, is then used to assign each node in our tree $T'^{(\mathbf{x})}$ to a particular group. Based on our assumptions that $\pi_i(\mathbf{x}) \equiv \pi_i$ and that the conditional distribution of $\mathbf{Y}$ belongs to a location-scale family, we have from the monotone invariance property (Proposition 4.1.2.1) that the set of feature variables chosen for $T'^{(\mathbf{x})}$ does not change depending on the value of $\mathbf{x}$. Using several examples, we have been able to construct $T'^{(\mathbf{x})}$ in the manner described in this section using MATLAB®.

4.2.3 Estimation of Unknown Conditional Distributions using Training Data

Suppose we do not have direct knowledge of the conditional CDFs $F_{\mathbf{Y}|\mathbf{x}}^{(i)}(\cdot)$ and only have access to the training data $(\mathbf{y}_{ij}, \mathbf{x}_{ij})$, the observed feature and covariate measurements for the j^{th} individual sampled from the i^{th} group $(i = 1, \ldots, g; j = 1, \ldots, n_i)$. As in the traditional case, the priors $\pi_i(\mathbf{x})$ may be specified in advance or estimated from the training data, which can be carried out using several estimation procedures, including logistic regression. With regards to $F_{\mathbf{Y}|\mathbf{x}}^{(1)}(\cdot), \ldots, F_{\mathbf{Y}|\mathbf{x}}^{(g)}(\cdot)$, we now describe two approaches we can take to estimate these g CDFs from the training data.

4.2.3.1 Parametric Approach We may assume that the conditional distribution of $\mathbf{Y}$ in the i^{th} group belongs to some known parametric family of distributions and use the training data to estimate all unknown parameters. Once we obtain the estimated parameters, we can use the assumed distribution of $\mathbf{Y}$ to obtain the estimated conditional probabilities $\hat{P}_{\mathbf{x}}^{(i)}(\mathbf{Y} \in t)$, $\hat{P}_{\mathbf{x}}^{(i)}(\mathbf{Y} \in t_L)$, and $\hat{P}_{\mathbf{x}}^{(i)}(\mathbf{Y} \in t_R)$. Using these probability estimates, along with the values of $\pi_i(\mathbf{x})$, our tree construction procedure can then proceed as in Section 4.1.2.1, terminating only when the number of observations in node t is less than some user defined value.

For example, we might assume that given $\mathbf{X} = \mathbf{x}$, $\mathbf{Y} \sim N_P(\boldsymbol{\mu}_{Y|X,i}, \boldsymbol{\Sigma}_{Y|X})$ in the i^{th} group and that $\mathbf{Y}$ has prior probability π_i of belonging to the i^{th} group. Based on the

training data, we can estimate $\boldsymbol{\mu}_{Y,i}$, $\boldsymbol{\Sigma}_{YX}$, $\boldsymbol{\Sigma}_{XX}$, $\boldsymbol{\mu}_X$, and $\boldsymbol{\Sigma}_{Y|X}$ using ML estimation. At this point, we note that the covariate adjusted feature vector $\tilde{\mathbf{Y}} \sim N_P(\boldsymbol{\mu}_{Y,i}, \boldsymbol{\Sigma}_{Y|X})$ in the i^{th} group, where $\tilde{\mathbf{Y}}$ is defined as in Section 3.2.2. We can then plug the ML estimates $\hat{\boldsymbol{\mu}}_{Y,i}$ and $\hat{\boldsymbol{\Sigma}}_{Y|X}$ into the P-variate normal density for $\tilde{\mathbf{Y}}$ in order to estimate the probabilities $P^{(i)}(\tilde{\mathbf{Y}} \in t)$, $P^{(i)}(\tilde{\mathbf{Y}} \in t_L)$, and $P^{(i)}(\tilde{\mathbf{Y}} \in t_R)$. Once we estimate these probabilities and obtain the values of π_i, we can use the population based method described in Section 4.2.2.2 to construct our covariate adjusted tree, which we can carry out using MATLAB®. Using the monotone invariance property (Proposition 4.1.2.1), along with our assumptions that the prior probabilities are not covariate dependent and that the conditional distribution of $\mathbf{Y}$ belongs to a location-scale family, we note that the same set of feature variables is chosen for a tree constructed using either the conditional distribution of $\mathbf{Y}$ or the distribution of $\tilde{\mathbf{Y}}$.

4.2.3.2 Non-parametric Approach On the other hand, we may have no knowledge regarding the conditional distribution functions $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ $(i = 1, \ldots, g)$, in which case we must obtain a non-parametric estimate of $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$. If we were, for a moment, to make the simplifying assumption that $\mathbf{X}$ takes on only discrete values, one possible estimate of $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ is

$$\hat{F}^{(i)}_{\mathbf{Y}|\mathbf{x}}(\mathbf{c}) = \hat{P}^{(i)}_{\mathbf{x}}(Y_1 \leq c_1, \ldots, Y_P \leq c_P) = \frac{\sum_{j=1}^{n_i} I(y_{ji,1} \leq c_1, \ldots, y_{ji,P} \leq c_P, \mathbf{x}_{ji} = \mathbf{x})}{\sum_{j=1}^{n_i} I(\mathbf{x}_{ji} = \mathbf{x})}, \quad (4.18)$$

which is applicable for both continuous and quantitative discrete feature data. Using the estimates $\hat{F}^{(i)}_{\mathbf{Y}|\mathbf{x}}(\mathbf{c})$ and the values of $\pi_i(\mathbf{x})$, we can construct a tree for each $\mathbf{x}$ value from the training feature data using the standard non-parametric approach described in Section 4.1.3.2. However, the estimated CDF in (4.18) does not make sense if $\mathbf{X}$ is continuous and creates a number of problems even if $\mathbf{X}$ is discrete, including the fact that a considerable amount of information is lost when computing the estimate in (4.18) because it is obtained from a comparatively small percentage of the training data [20].

Li and Racine[20] and Peracchi[28] have developed procedures to estimate the conditional CDFs $F^{(i)}_{Y|\mathbf{x}}(\centerdot)$ $(i = 1, \ldots, g)$ for univariate $\mathbf{Y}$ using kernel density estimation and semi-parametric estimation, respectively. If we have univariate feature data, we can use their estimates of $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ and our tree construction procedure can proceed as in Section 4.1.2.1.

It is clear that if we have no knowledge regarding the conditional distribution of $\mathbf{Y}$, then there appears to be no available method of computing the empirical conditional CDFs for the feature data in a way that is generally applicable for all $\mathbf{Y}$ and $\mathbf{X}$. More importantly, unless certain conditions hold for the conditional distribution of the feature data, the feature variables chosen for the conditional tree $T'^{(\mathbf{x})}$ constructed using any of the estimates of $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ described in Section 4.2.3 changes depending on the value of $\mathbf{x}$, which may create interpretation problems in certain contexts. To address these two key concerns that arise in our construction of $T'^{(\mathbf{x})}$, we develop certain conditions for the conditional distribution of $\mathbf{Y}$ for a given $\mathbf{x}$ so that the feature variables chosen for $T'^{(\mathbf{x})}$ do not depend on $\mathbf{x}$ and the conditional CDFs $F^{(i)}_{Y|\mathbf{x}}(\centerdot)$ can be estimated empirically in a way that can be handled by the standard BFOS tree construction algorithm. This is the topic of the next section.

4.3 SEMI-PARAMETRIC CLASSIFICATION TREES

4.3.1 Motivation

As we previously pointed out, there are two important issues that arise, a primary one and a secondary one, when we condition on $\mathbf{X} = \mathbf{x}$ in our construction of a particular classification tree.

The primary issue is the fact that if we have no knowledge regarding the conditional distribution of $\mathbf{Y}$ in each group, then there is no completely non-parametric method of estimating the conditional distribution functions $F^{(1)}_{\mathbf{Y}|\mathbf{x}}(\centerdot), \ldots, F^{(g)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ for a given $\mathbf{x}$ that applies regardless of whether $\mathbf{X}$ is discrete or continuous or whether $\mathbf{Y}$ is univariate or multivariate. Therefore, our primary goal is to extend the standard non-parametric procedure in Section 4.1.3.2 so that it is applicable for all continuous $\mathbf{Y}$ and all $\mathbf{X}$ in general. We note that Lachenbruch and Tu et al. developed general covariance adjusted LDA by simply applying traditional LDA to feature data from which all effects of some known function of $\mathbf{x}$ were removed, i.e., feature data adjusted for all relevant covariate effects. Although various assumptions were made in their development, the assumption that the known function of $\mathbf{x}$ did not depend on group mainly contributed to the fact that Lachenbruch and Tu et al. were

able to work with the adjusted feature data while retaining all other aspects of traditional LDA.

Thus, if we could construct a tree for a given $\mathbf{x}$ using an estimate of the conditional CDF in the i^{th} group based on the covariate adjusted feature vector, we can still account for the effects of $\mathbf{x}$ without having to worry about the number of feature variables in $\mathbf{Y}$ or whether $\mathbf{X}$ is discrete or continuous. Therefore, our primary goal is to develop a model for the conditional distribution of $\mathbf{Y}$ for a given $\mathbf{x}$ in each group that allows us to construct a conditional classification tree based on training data by implementing the standard BFOS non-parametric tree construction procedure on training feature data that have been adjusted for all relevant covariate effects.

The secondary issue that arises when we condition on $\mathbf{X} = \mathbf{x}$ in our tree construction procedure is the fact that the feature variable Y in $\mathbf{Y}$ and the cutpoint c chosen to split a particular node t into t_L and t_R both depend on the value at which $\mathbf{x}$ is fixed. For example, if we condition on gender, then the feature variable and cutpoint that are chosen to split node t may depend on whether $\mathbf{Y}$ corresponds to a male or female individual. Although there is nothing intrinsically wrong with having a different set of optimal splitting variables depending on a particular value of $\mathbf{x}$, it does not make sense in certain contexts. For example, in the context of post-mortem tissue studies, having a subset of biomarkers that best discriminates between the control and schizophrenia diagnostic groups depend on tissue storage time or PMI may be of little practical use because one would not expect tissue storage time or PMI to differentially affect which biomarkers are chosen for a particular tree. Conceptually, however, it is possible for the effect of subject age on the biomarker data to be differentially expressed and in such cases, the models we use in our formulation of conditional classification trees may need to be modified.

The goal of post-mortem tissue studies is to obtain a subset of discriminatory biomarkers, when appropriate, that does not vary depending on the value of the experimental covariate(s) we wish to account for. Therefore, in addition to our primary goal, we would also like to develop a model for the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ in each group such that regardless of the value that $\mathbf{x}$ takes on, only one subset of feature variables is chosen when constructing our conditional classification tree.

4.3.1.1 Linear Invariance Property

Proposition 4.3.1.1. *Suppose* $\mathbf{Y}$ *has CDF* $F^{(i)}_{\mathbf{Y}}(\cdot)$ *in the* i^{th} *group* $(i = 1, \ldots, g)$*. Given* $\mathbf{X} = \mathbf{x}$*, let* $\mathbf{Y}_{\mathbf{x}}$ *denote the translated feature vector with conditional distribution function* $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\cdot)$ *in the* i^{th} *group, such that* $\mathbf{Y}_{\mathbf{x}}$ *is equal in distribution to* $\mathbf{Y}+\boldsymbol{\xi}(\mathbf{x})$ *(i.e.,* $\mathbf{Y}_{\mathbf{x}} \stackrel{d}{=} \mathbf{Y}+\boldsymbol{\xi}(\mathbf{x})$*) and* $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1(\mathbf{x}), \ldots, \xi_P(\mathbf{x}))'$ *is a known function of* $\mathbf{x}$*. In addition, suppose* $\mathbf{Y}_{\mathbf{x}}$ *has prior probability* π_i *of belonging to the* i^{th} *group, regardless of the value of* $\mathbf{x}$*. Based on either the impurity measure based GOS criterion or the twoing criterion, let* $T'^{(\mathbf{x}_a)}$ *be the classification tree based on* $\pi_1, \ldots, \pi_g$ *and* $F^{(1)}_{\mathbf{Y}|\mathbf{x}_a}(\cdot), \ldots, F^{(g)}_{\mathbf{Y}|\mathbf{x}_a}(\cdot)$ *for covariate value* $\mathbf{x}_a$*, and* $T'^{(\mathbf{x}_b)}$ *be the classification tree based on* $\pi_1, \ldots, \pi_g$ *and* $F^{(1)}_{\mathbf{Y}|\mathbf{x}_b}(\cdot), \ldots, F^{(g)}_{\mathbf{Y}|\mathbf{x}_b}(\cdot)$ *for covariate value* $\mathbf{x}_b$*. Then,* $T'^{(\mathbf{x}_a)}$ *and* $T'^{(\mathbf{x}_b)}$ *have the same set of splitting variables and the set of cutpoints for* $T'^{(\mathbf{x}_a)}$*,* $\mathbf{c}_{T'^{(\mathbf{x}_a)}}$*, are related to those of* $T'^{(\mathbf{x}_b)}$*,* $\mathbf{c}_{T'^{(\mathbf{x}_b)}}$*, by* $\mathbf{c}_{T'^{(\mathbf{x}_b)}} = \mathbf{c}_{T'^{(\mathbf{x}_a)}} - \boldsymbol{\xi}(\mathbf{x}_a) + \boldsymbol{\xi}(\mathbf{x}_b)$*.*

Proof. For a given $\mathbf{x}_a$ and $\mathbf{x}_b$, $\mathbf{Y}_{\mathbf{x}_b} = \mathbf{Y}_{\mathbf{x}_a} - \boldsymbol{\xi}(\mathbf{x}_a) + \boldsymbol{\xi}(\mathbf{x}_b)$. In other words, $\mathbf{Y}_{\mathbf{x}_b} = \boldsymbol{\zeta}(\mathbf{Y}_{\mathbf{x}_a})$ is an increasing linear function of $\mathbf{Y}_{\mathbf{x}_a}$, i.e., a monotonic transformation of $\mathbf{Y}_{\mathbf{x}_a}$, where $\boldsymbol{\zeta}(\mathbf{Y}) = \mathbf{Y} - \boldsymbol{\xi}(\mathbf{x}_a) + \boldsymbol{\xi}(\mathbf{x}_b)$. It now directly follows from the monotone invariance property (Proposition 4.1.2.1) that the set of splitting variables for $T'^{(\mathbf{x}_a)}$ and $T'^{(\mathbf{x}_b)}$ are the same and that $\mathbf{c}_{T'^{(\mathbf{x}_b)}} = \mathbf{c}_{T'^{(\mathbf{x}_a)}} - \boldsymbol{\xi}(\mathbf{x}_a) + \boldsymbol{\xi}(\mathbf{x}_b)$. □

Therefore, if the prior probability of group membership for $\mathbf{Y}$ does not depend on $\mathbf{x}$ and the conditional distribution of $\mathbf{Y}$ for a given $\mathbf{x}$ is simply a location shift of the distribution of $\mathbf{Y}$ by the known function $\boldsymbol{\xi}(\mathbf{x})$, then we can be assured that the set of splitting variables chosen for our conditional tree $T'^{(\mathbf{x})}$ will not change depending on $\mathbf{x}$, as we pointed out in Section 4.2.2.1. From the linear invariance property, we see that if $\boldsymbol{\xi}(\mathbf{x})$ depended on group, then the set of cutpoints for a particular tree would also depend on group, which does not make sense in the context of classification trees. Thus, it is necessary and reasonable to assume that the function $\boldsymbol{\xi}(\mathbf{x})$ does not depend on group, an assumption Tu et al. also make for general covariance adjusted LDA.

Proposition 4.3.1.1 lays the groundwork for a model for the conditional distribution of $\mathbf{Y}$ for a given value of $\mathbf{x}$ to ensure that the feature variable chosen to split a particular node t in $T'^{(\mathbf{x})}$ does not depend on $\mathbf{x}$. In addition, we show how this model allows us to implement the standard BFOS tree construction procedure on feature data that have been suitably adjusted for covariate effects.

4.3.2 Proposed Model for Known Conditional Distributions

Given $\mathbf{X} = \mathbf{x}$, we let $\mathbf{Y_x}$ denote the random feature vector conditional on $\mathbf{x}$ with known CDF $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\mathbf{c}) = F^{(i)}_{\mathbf{Y}}(\mathbf{c} - \boldsymbol{\xi}(\mathbf{x}; \boldsymbol{\Theta}))$ in the i^{th} group ($i = 1, \ldots, g$), where $F^{(i)}_{\mathbf{Y}}(\centerdot)$ are fixed CDFs, i.e., they do not depend on $\mathbf{x}$, and $\boldsymbol{\xi}(\mathbf{x}; \boldsymbol{\Theta}) = (\xi_1(\mathbf{x}|\boldsymbol{\theta}_1), \ldots, \xi_P(\mathbf{x}|\boldsymbol{\theta}_P))'$ is a known smooth function of $\mathbf{x}$ and known parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$. Thus, if $\mathbf{Y}$ has distribution function $F^{(i)}_{\mathbf{Y}}(\centerdot)$ in the i^{th} group, then $\mathbf{Y_x} \stackrel{d}{=} \mathbf{Y} + \boldsymbol{\xi}(\mathbf{x}; \boldsymbol{\Theta})$. In addition, we let $\mathbf{Y_x}$ have prior probability π_i of belonging to the i^{th} group, regardless of $\mathbf{x}$. Our model for the conditional CDFs $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ can be viewed as semi-parametric because it has both a parametric component, namely, the parametric function $\boldsymbol{\xi}(\mathbf{x}; \boldsymbol{\Theta})$, and the non-parametric component $F^{(i)}_{\mathbf{Y}}(\centerdot)$ ($i = 1, \ldots, g$).

Alternatively, we may examine the covariate adjusted feature vector $\tilde{\mathbf{Y}} = \mathbf{Y_x} - \boldsymbol{\xi}(\mathbf{x}; \boldsymbol{\Theta})$ with prior probability π_i and known CDF $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$ in the i^{th} group, where $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot) \equiv F^{(i)}_{\mathbf{Y}}(\centerdot)$. Since the CDFs $F^{(i)}_{\mathbf{Y}}(\centerdot)$ and $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$ are the same, the conditional probabilities $P^{(i)}_{\mathbf{x}}(\mathbf{Y_x} \in t)$, $P^{(i)}_{\mathbf{x}}(\mathbf{Y_x} \in t_L)$, and $P^{(i)}_{\mathbf{x}}(\mathbf{Y_x} \in t_R)$ obtained from $F^{(i)}_{\mathbf{Y}|\mathbf{x}}(\centerdot)$ can equivalently be expressed as $P^{(i)}(\tilde{\mathbf{Y}} \in t)$, $P^{(i)}(\tilde{\mathbf{Y}} \in t_L)$, and $P^{(i)}(\tilde{\mathbf{Y}} \in t_R)$ obtained from $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$.

Assuming equal misclassification costs, we can construct the covariate adjusted tree $T'^{\mathrm{adj}(\mathbf{x})}$ using the traditional population-based method in Section 4.1.2.1 by simply replacing the probabilities $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ used to construct T' with the probabilities $P^{(i)}(\tilde{\mathbf{Y}} \in t)$, $P^{(i)}(\tilde{\mathbf{Y}} \in t_L)$, and $P^{(i)}(\tilde{\mathbf{Y}} \in t_R)$. The tree $T'^{\mathrm{adj}(\mathbf{x})}$ can then be used to classify a randomly selected individual in the population into one of the g groups based on this individual's covariate adjusted feature measurement $\tilde{\mathbf{y}}$.

Therefore, in assuming that our feature vector $\mathbf{Y}$ is shifted by the known function $\boldsymbol{\xi}(\mathbf{x}; \boldsymbol{\Theta})$, we can implement the population-based BFOS recursive partitioning algorithm on the covariate adjusted feature vector $\tilde{\mathbf{Y}}$ in the population setting to construct the tree $T'^{\mathrm{adj}(\mathbf{x})}$ that suitably adjusts for the effects of the covariate vector $\mathbf{X}$. Based on the linear invariance property (Proposition 4.3.1.1), the following facts regarding $T'^{\mathrm{adj}(\mathbf{x})}$ hold: (1) regardless of the value of $\mathbf{x}$, $T'^{\mathrm{adj}(\mathbf{x})}$ helps us identify a unique set of feature variables that best discriminates among the g groups under consideration while accounting for all relevant covariate effects; (2) if $\tilde{Y}_\nu$ and $Y_{\nu,\mathbf{x}}$ correspond to any of the P feature variables in $\tilde{\mathbf{Y}}$ and $\mathbf{Y_x}$ ($\nu \in (1, 2, \ldots, P)$), respectively, then the split $\tilde{Y}_\nu \leq \tilde{c}_\nu$ in $T'^{\mathrm{adj}(\mathbf{x})}$ is equivalent to the split $Y_{\nu,\mathbf{x}} \leq \tilde{c}_\nu + \xi_\nu(\mathbf{x}|\boldsymbol{\theta}_\nu)$ in the tree based on $\mathbf{Y_x}$. For example, in the context of post-mortem studies, if a biomarker

adjusted for the effect of storage time is greater than a constant value for a specific split in our adjusted tree, then this biomarker, when expressed on the original scale, will be greater than a value that will depend on storage time.

4.3.3 Tree Construction for the Semi-Parametric Model, using Training Data

Our first step in estimating $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$ $(i = 1, \ldots, g)$ is to estimate the parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ from the training data $(\mathbf{y}_{ij}, \mathbf{x}_{ij})$, after which we can estimate the CDF of $\tilde{\mathbf{Y}}$ in each of the g groups. In the next two sections, we provide a method that can be used to estimate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$, and follow with a discussion of how to estimate $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$ based on the estimates of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$.

4.3.3.1 Estimation of Unknown Parameters We begin by first estimating the parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ from the training data. A simple approach we use is to assume that the conditional mean of $\mathbf{Y}_{ij} = (Y_{ij,1}, \ldots, Y_{ij,P})'$, the random feature vector corresponding to the j^{th} individual randomly sampled from the i^{th} group $(i = 1, \ldots, g;\ j = 1, \ldots, n_i)$, is given by $E[\mathbf{Y}_{ij}|\mathbf{x}_{ij}] = \boldsymbol{\lambda}_i + \boldsymbol{\xi}(\mathbf{x}_{ij}; \boldsymbol{\Theta}) = (\lambda_{1,i} + \xi_1(\mathbf{x}_{ij}|\boldsymbol{\theta}_1), \ldots, \lambda_{P,i} + \xi_P(\mathbf{x}_{ij}|\boldsymbol{\theta}_P))'$ (which uses Lachenbruch and Tu et al.'s notation), where $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \ldots, \lambda_{P,i})'$ corresponds to the effect of $\mathbf{Y}_{ij}$ belonging to the i^{th} group. We can then use least squares (LS) estimation to obtain the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$, i.e., $\hat{\boldsymbol{\theta}}_p$ is the value of $\boldsymbol{\theta}_p$ that minimizes the LS criterion $Q_p = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij,p} - \lambda_{p,i} - \xi_p(\mathbf{x}_{ij}|\boldsymbol{\theta}_p))^2$ $(p = 1, \ldots, P)$.

4.3.3.2 Tree Construction Procedure Once we obtain the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$, we can easily compute the covariate adjusted training feature data $\hat{\tilde{\mathbf{y}}}_{ij} = (\hat{\tilde{y}}_{ij,1}, \ldots, \hat{\tilde{y}}_{ij,P})' = (y_{ij,1} - \xi_1(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_1), \ldots, y_{ij,P} - \xi_P(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_P))' = \mathbf{y}_{ij} - \boldsymbol{\xi}(\mathbf{x}_{ij}; \hat{\boldsymbol{\Theta}})$. Holding $\mathbf{x}_{ij}$, $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ fixed, we then view $\hat{\tilde{\mathbf{y}}}_{ij}$ as a random sample of n_i observations from $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$, so that we can estimate $F^{(i)}_{\tilde{\mathbf{Y}}}(\centerdot)$ as

$$\hat{F}^{(i)}_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{c}}) = \hat{P}^{(i)}(\tilde{Y}_1 \leq \tilde{c}_1, \ldots, \tilde{Y}_P \leq \tilde{c}_P) = \frac{\sum_{j=1}^{n_i} I(\hat{\tilde{y}}_{ij,1} \leq \tilde{c}_1, \ldots, \hat{\tilde{y}}_{ij,P} \leq \tilde{c}_P)}{n_i}, \tag{4.19}$$

and estimate $P^{(i)}(\tilde{\mathbf{Y}} \in t)$ as

$$\hat{P}^{(i)}(\tilde{\mathbf{Y}} \in t) = \frac{\sum_{j=1}^{n_i} I(\hat{\tilde{\mathbf{y}}}_{ij} \in t)}{n_i}, \tag{4.20}$$

the sample proportion of covariate adjusted feature observations in group i that fall in node t. Based on π_i and the probability estimates in (4.20), we can implement the traditional non-parametric approach in Section 4.1.3.2 on the adjusted training feature data to construct our adjusted tree and prune it accordingly using the minimal cost-complexity pruning procedure described in Section 4.1.3.4. Our construction of a covariate adjusted tree in this manner can be easily achieved using standard software packages such as R or Salford Systems CART®.

As was the case in the population setting, we can apply the linear invariance property (Proposition 4.3.1.1) to state that the cutpoints for our covariate adjusted tree are fixed constants when expressed in terms of the adjusted feature variables, but are covariate dependent when expressed in terms of the original feature variables, even though the directionality of a particular tree split is preserved for both the adjusted and original feature data.

Depending on the structure of $\xi_p(\mathbf{x};\boldsymbol{\theta}_p)$ $(p = 1,\ldots,P)$, the parameter vectors $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_P$ may not be identifiable, in which case the LS estimates $\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_P$ are not unique. However, regardless of the values of $\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_P$, our covariate adjusted tree yields the same discrimination and classification results. Specifically, suppose we obtain two LS estimates of $\boldsymbol{\Theta}$, namely, $\hat{\boldsymbol{\Theta}}_a$ and $\hat{\boldsymbol{\Theta}}_b$ $(\hat{\boldsymbol{\Theta}}_a \neq \hat{\boldsymbol{\Theta}}_b)$, such that $\boldsymbol{\xi}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}}_a) = (\xi_1(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_{1,a}),\ldots,\xi_P(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_{P,a}))'$ and $\boldsymbol{\xi}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}}_b) = (\xi_1(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_{1,b}),\ldots,\xi_P(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_{P,b}))'$. Consider the covariate adjusted trees $T_a^{'\mathrm{adj}(\mathbf{x})}$ and $T_b^{'\mathrm{adj}(\mathbf{x})}$ constructed using the two adjusted data sets $\hat{\tilde{\mathbf{y}}}_{ij,a} = \mathbf{y}_{ij} - \boldsymbol{\xi}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}}_a)$ and $\hat{\tilde{\mathbf{y}}}_{ij,b} = \mathbf{y}_{ij} - \boldsymbol{\xi}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}}_b)$, respectively, where $\hat{\tilde{\mathbf{y}}}_{ij,b} = \hat{\tilde{\mathbf{y}}}_{ij,a} + \boldsymbol{\xi}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}}_a) - \boldsymbol{\xi}(\mathbf{x}_{ij};\hat{\boldsymbol{\Theta}}_b)$. Based on the linear invariance property (Proposition 4.3.1.1), the following facts hold: (1) the same set of feature variables is chosen for $T_a^{'\mathrm{adj}(\mathbf{x})}$ and $T_b^{'\mathrm{adj}(\mathbf{x})}$; (2) the split $\tilde{Y}_{\nu,a} \leq \tilde{c}_{\nu,a}$ in $T_a^{'\mathrm{adj}(\mathbf{x})}$ $(\nu \in (1,2,\ldots,P))$ is equivalent to the split $\tilde{Y}_{\nu,b} \leq \tilde{c}_{\nu,a} + \xi_\nu(\mathbf{x}|\hat{\boldsymbol{\theta}}_{\nu,a}) - \xi_\nu(\mathbf{x}|\hat{\boldsymbol{\theta}}_{\nu,b})$ in $T_b^{'\mathrm{adj}(\mathbf{x})}$ for any given covariate value $\mathbf{x}$. In other words, the observations that fall in the left descendant node of the split $\tilde{Y}_{\nu,a} \leq \tilde{c}_{\nu,a}$ are identical to those that fall in the left descendant node of the split $\tilde{Y}_{\nu,b} \leq \tilde{c}_{\nu,a} + \xi_\nu(\mathbf{x}|\boldsymbol{\theta}_{\nu,a}) - \xi_\nu(\mathbf{x}|\boldsymbol{\theta}_{\nu,b})$ and likewise for the right descendant nodes, so that $T_a^{'\mathrm{adj}(\mathbf{x})}$ and $T_b^{'\mathrm{adj}(\mathbf{x})}$ yield the same classification results.

In short, when we construct a tree based on the covariate adjusted training feature data, we obtain a tree that helps us determine the set of feature variables and corresponding splits that best discriminates among the g groups of interest, while, at the same time, accounting for covariate effects on the feature data.

4.3.4 Summary of Semi-Parametric Classification Trees

Our formulation of semi-parametric classification trees has a number of desirable properties. First, and perhaps most importantly, it allows us to use the traditional BFOS recursive partitioning algorithm in both the population and data settings to construct a tree based on the covariate adjusted feature vector $\tilde{\mathbf{Y}}$ that adjusts for the effects of the covariate vector $\mathbf{X}$. In particular, we can construct such a tree using available training data and prune it accordingly using any software package that implements the standard non-parametric approach described in Section 4.1.3.2 and minimal cost-complexity pruning as described in Section 4.1.3.4. Therefore, if we wish to construct semi-parametric classification trees in the data setting, there is no need to develop new software packages to do so.

In addition, the development of our semi-parametric conditional model for the feature data helps us obtain a covariate adjusted tree that allows us to not only classify new individuals, but also identify a unique set of feature variables and corresponding splits that best discriminates among the g groups of interest, while accounting for all relevant covariate effects. For example, in the context of post-mortem tissue studies, semi-parametric classification trees can help us identify which biomarkers best discriminate between the control and schizophrenia diagnostic groups, without the confounding effects of additional covariates, e.g., brain tissue storage time.

We now discuss two extensions of our semi-parametric tree construction methodology to handle the case where individuals are matched across two or more groups and measured on additional covariates. In Section 4.4, we develop a methodology to adjust for the effect of group matching on the feature variables of interest and then extend our matched adjustment methodology in Section 4.5 to also account for covariate effects.

4.4 MATCHED CLASSIFICATION TREES

4.4.1 Known Distributions

In this section, our focus is on constructing a classification tree that accounts for the effect of subject matching on the feature data, so that we may more accurately determine the subset of feature variables and corresponding splits that best discriminates among the g

($g \geq 2$) groups under consideration, as well as classify each individual belonging to a new g-tuple or match. For example, with regards to the Konopaske biomarker data, we would like to account for the effect of triad matching on the examined biomarkers in order to better identify the biomarkers that best discriminate among the haloperidol, olanzapine, and sham treatment groups. In addition, in the context of post-mortem tissue studies where normal controls and schizophrenia subjects are paired on certain characteristics, we want to adjust for the effect of subject pairing on the biomarker data when constructing our tree, so that we may have a clearer picture of which biomarkers and corresponding splits best distinguish a normal control from an individual with schizophrenia subject in a given pair.

First, we consider the conditional distribution of $\mathbf{Y}$ for a given match. As we did when we adjusted for subject matching in LDA, we let the parameter vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_P)'$ correspond to each individual in a match across the P feature variables, where $\boldsymbol{\gamma}$ denotes the effect of group matching on $\mathbf{Y}$. Using our semi-parametric conditional model based on the estimated parameter vectors $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ from the training data, we compute the covariate adjusted feature data for a new individual whose measurement is not part of the training data, and classify this individual using his or her covariate adjusted feature measurement. When we include $\boldsymbol{\gamma}$ in our conditional model to account for matching, however, we must re-estimate $\boldsymbol{\gamma}$ for each new individual in a match (i.e., an individual not included in the training data), since $\boldsymbol{\gamma}$ is specific to each match. The procedure we develop to account for the effect of group matching on the feature data can be viewed as an extension of our semi-parametric model to include parameters that must be re-estimated for each individual in a match that is not part of the training data, along with parameters that are estimated solely from available training data.

We first present the general case where individuals are matched across two or more groups, where we develop three different approaches to account for the effect of group matching from a population based perspective. When given a set of feature measurements for all g members of a match, we know that the first member belongs to group i_1, the second member belongs to group i_2, ..., and the g^{th} member belongs to group i_g ($i_1, i_2, \ldots, i_g = 1, \ldots, g$; $i_1 \neq i_2 \neq \cdots \neq i_g$). In this case, it is equally likely that each member belongs to one of the g groups, i.e., $\pi_i = 1/g$ ($i = 1, \ldots, g$), since we assume there is no preference for which member is labeled first, second, etc. We also retain our assumption

in Section 4.3.1.1 of equal misclassification costs.

4.4.1.1 Tree Construction using Feature Vector, adjusting for Effect of Matching For known $\boldsymbol{\gamma}$, we let $\mathbf{Y}_{\boldsymbol{\gamma},ind}$ denote the random feature vector for any individual belonging to a particular match and let $\mathbf{Y}_{\boldsymbol{\gamma},ind}$ have known CDF $F^{(i)}_{\mathbf{Y}_{ind}}(\mathbf{c}-\boldsymbol{\gamma})$ in the i^{th} group $(i=1,\ldots,g)$, where $\mathbf{Y}_{ind}$ denotes the random feature vector for any individual in a match and has CDF $F^{(i)}_{\mathbf{Y}_{ind}}(\mathbf{c})$ in the i^{th} group, i.e., $\mathbf{Y}_{\boldsymbol{\gamma},ind} \stackrel{d}{=} \mathbf{Y}_{ind}+\boldsymbol{\gamma}$.

We may also examine the feature vector for an individual that has been suitably adjusted for the effect of matching, namely, $\tilde{\mathbf{Y}}_{ind} = \mathbf{Y}_{\boldsymbol{\gamma},ind}-\boldsymbol{\gamma}$ with known CDF $F^{(i)}_{\tilde{\mathbf{Y}}_{ind}}(\centerdot)$ in the i^{th} group. Since $F^{(i)}_{\tilde{\mathbf{Y}}_{ind}}(\centerdot) \equiv F^{(i)}_{\mathbf{Y}_{ind}}(\centerdot)$, the probabilities obtained from $F^{(i)}_{\mathbf{Y}_{ind}}(\mathbf{c}-\boldsymbol{\gamma})$ can equivalently be obtained from $F^{(i)}_{\tilde{\mathbf{Y}}_{ind}}(\centerdot)$.

Under our assumptions of equal priors and equal misclassification costs, we can construct a tree $T'^{\mathrm{adj}(\boldsymbol{\gamma})}$ that adjusts for the effect of matching on the feature data by using the traditional population-based approach in Section 4.1.2.1, replacing the probabilities $P^{(i)}(\mathbf{Y}\in t)$, $P^{(i)}(\mathbf{Y}\in t_L)$, and $P^{(i)}(\mathbf{Y}\in t_R)$ used to construct T' with the probabilities $P^{(i)}(\tilde{\mathbf{Y}}_{ind}\in t)$, $P^{(i)}(\tilde{\mathbf{Y}}_{ind}\in t_L)$, and $P^{(i)}(\tilde{\mathbf{Y}}_{ind}\in t_R)$ obtained from $F^{(i)}_{\tilde{\mathbf{Y}}_{ind}}(\centerdot)$. The adjusted tree $T'^{\mathrm{adj}(\boldsymbol{\gamma})}$ can then be used to classify any individual in a match based on their adjusted feature measurement $\tilde{\mathbf{y}}_{ind}$. More importantly, we can use $T'^{\mathrm{adj}(\boldsymbol{\gamma})}$ to identify the set of feature variables, once the effect of group matching has been adjusted for, and corresponding splits that best discriminate among the g groups. In the context of post-mortem tissue studies, this adjusted tree can be used to identify the biomarkers and splits on these biomarkers that best discriminate between the control and schizophrenia diagnostic groups, once we adjust for the effect of subject pairing on the biomarker data. With regards to the Konopaske et al. data, the tree $T'^{\mathrm{adj}(\boldsymbol{\gamma})}$ can help us determine which biomarkers, suitably adjusted for the effect of triad matching, and corresponding splits best differentiate among the haloperidol, olanzapine, and sham treatment groups.

4.4.1.2 Tree Construction using Differenced Feature Vector An alternate method we develop to account for the effect of group matching when constructing a classification tree is to apply the traditional BFOS recursive partitioning algorithm to the differenced random feature vector $\mathbf{Y}_{\mathrm{diff}} \equiv \mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}$, where $\mathbf{Y}_{ind}, \mathbf{Y}_{sib,1},\ldots,\mathbf{Y}_{sib,g-1}$ denote the

random feature vectors for any individual and their $g-1$ siblings in a given match.

To clarify, we begin with the assumption that $\mathbf{Y}_{\text{diff}}$ has known CDF

$$F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\mathbf{c}) = P^{(i)}(Y_{1,\text{diff}} \leq c_1, \ldots, Y_{P,\text{diff}} \leq c_P) \tag{4.21}$$

in the i^{th} population ($i = 1, \ldots, g$), where $\mathbf{Y}_{ind}$ belongs to the i^{th} group. For the same reasons as those stated in Section 3.5.1.2, we assume that the prior probability of each population of $\mathbf{Y}_{\text{diff}}$ is equal to $1/g$.

In this case, the tree $T'^{(\text{diff})}$ can be constructed using the standard approach in Section 4.1.2.1 by replacing $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ used to construct T' with $P^{(i)}(\mathbf{Y}_{\text{diff}} \in t)$, $P^{(i)}(\mathbf{Y}_{\text{diff}} \in t_L)$, and $P^{(i)}(\mathbf{Y}_{\text{diff}} \in t_R)$ based on the CDFs $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\centerdot)$, where $P^{(i)}(\mathbf{Y}_{\text{diff}} \in t) = P(\mathbf{Y}_{\text{diff}} \in t | \mathbf{Y}_{\text{diff}} \in \text{population } i)$. We then use the following rule to assign each terminal node t of $T'^{(\text{diff})}$ to the i^{th} population of $\mathbf{Y}_{\text{diff}}$, based on our assumption of equal priors and equal misclassification costs:

$$R^{\text{diff}}_i : \left\{ t : P^{(i)}(\mathbf{Y}_{\text{diff}} \in t) > P^{(j)}(\mathbf{Y}_{\text{diff}} \in t) \right\}, \quad j = 1, \ldots, g;\ j \neq i. \tag{4.22}$$

For each individual in a match, we compute the difference $\mathbf{y}_{\text{diff}} = \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{y}_{sib,m}$, the difference between the feature measurement for that individual and the average of the feature measurements for their siblings in that match. If this difference falls into a terminal node of $T'^{(\text{diff})}$ that has been assigned to the i^{th} population according to the rule in (4.22), then we classify that individual into the i^{th} group ($i = 1, \ldots, g$).

When we're dealing with matched pairs, one notable difference exists between the classification results obtained from our pairwise differencing approaches for LDA and classification trees in Sections 3.3.1.2 and 4.4.1.2, respectively. Recall that in our discussion of paired LDA, the classification regions in (3.16) based on the pairwise difference $\mathbf{y}_{ind} - \mathbf{y}_{sib}$ were obtained by comparing the linear discriminant function $(\mathbf{y}_{ind} - \mathbf{y}_{sib})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ to the cutpoint of zero, which ensured that an individual and their sibling in a pair would be classified into different groups. On the other hand, if the tree $T'^{(\text{diff})}$ is constructed based on the pairwise difference $\mathbf{Y}_{ind} - \mathbf{Y}_{sib} \equiv \mathbf{Y}_{\text{diff}}$, it is possible that $T'^{(\text{diff})}$ will classify an individual and their sibling in a pair into the same group. For example, in our construction of $T'^{(\text{diff})}$, suppose we only split once on one of the P elements of $\mathbf{Y}_{\text{diff}}$, namely, $Y_{\text{diff},\nu}$ ($\nu \in (1, 2, \ldots, P)$) whose corresponding optimal cutpoint c_ν is some positive value, so that the left and right terminal

nodes of $T'^{(\text{diff})}$ are assigned to the first and second populations, respectively, based on the rule in (4.22). If the value of $y_{\text{diff},\nu}$ satisfies $0 < y_{\text{diff},\nu} \leq c_\nu$, then $y_{\text{diff},\nu}$ falls into the left terminal node of $T'^{(\text{diff})}$ and we would classify this individual into the first group. However, since this individual's sibling is classified based on the difference $-y_{\text{diff},\nu}$, where in this example $-y_{\text{diff},\nu} < 0 < c_\nu$, we would also classify this sibling into the first group. In addition, we note that it also possible for the tree $T'^{\text{adj}(\gamma)}$ in Section 4.4.1.1 to classify both members of a pair into the same group.

In conclusion, we can use $T'^{(\text{diff})}$ to identify among the feature variables of interest those that best discriminate among the g groups, once we account for the effect of group matching on these feature variables. In the case of subject pairing, we note that $T'^{(\text{diff})}$ can be used to determine which feature variables best distinguish an individual belonging to group 1 from that belonging to group 2 in any given pair. In addition, the cutpoints of $T'^{(\text{diff})}$ can be used to determine whether large or small values of each splitting variable in $T'^{(\text{diff})}$ for an individual in a pair, relative to the values of the same splitting variable for the individual's sibling in the same pair, are associated with group 1 compared with group 2. For example, consider the case where we construct $T'^{(\text{diff})}$ based on biomarker data obtained from schizophrenia subjects that are paired with normal controls. Suppose that in our construction of $T'^{(\text{diff})}$, we only split once on one of the P differenced biomarkers in $\mathbf{Y}_{\text{diff}}$, so that the left and right terminal nodes of $T'^{(\text{diff})}$ are assigned to the control and schizophrenia populations, respectively, based on the rule in (4.22). In this case, we can infer from this tree that for any given pair, normal controls have smaller values of this biomarker relative to individuals with schizophrenia.

Intriguingly, when applied to matched data, the two adjustment approaches we develop in Sections 4.4.1.1 and 4.4.1.2 based on the adjusted random feature vector $\tilde{\mathbf{Y}}_{ind}$ and the differenced random feature vector $\mathbf{Y}_{\text{diff}}$, respectively, produce trees that have the same structure and identical sets of splitting variables, which we show in Section 4.4.2.2.

4.4.1.3 Tree Construction using Stacked Feature Vector, adjusting for Effect of Matching

Recall that for LDA, we developed a methodology to adjust for the effect of group matching based on the stacked random feature vector $\mathbf{Y}^+$. We now briefly discuss how this methodology can be implemented in the context of classification trees, and why it

does not yield practical results.

Clearly, we can apply our adjustment methodology in Section 4.4.1.1 to $\mathbf{Y}_{\boldsymbol{\gamma}}^{+} = \begin{bmatrix} \mathbf{Y}_{\boldsymbol{\gamma},ind} \\ \mathbf{Y}_{\boldsymbol{\gamma},sib,1} \\ \vdots \\ \mathbf{Y}_{\boldsymbol{\gamma},sib,g-1} \end{bmatrix}$, which denotes the random feature vector corresponding to an individual and their $g-1$ siblings in a given match for known $\boldsymbol{\gamma}$, where $\mathbf{Y}_{\boldsymbol{\gamma},ind}$ has known CDF $F_{\mathbf{Y}}^{(i_1)}(\mathbf{c}_{ind} - \boldsymbol{\gamma})$ in group i_1, $\mathbf{Y}_{\boldsymbol{\gamma},sib,1}$ has known CDF $F_{\mathbf{Y}}^{(i_2)}(\mathbf{c}_{sib,1} - \boldsymbol{\gamma})$ in group i_2, ..., and $\mathbf{Y}_{\boldsymbol{\gamma},sib,g-1}$ has known CDF $F_{\mathbf{Y}}^{(i_g)}(\mathbf{c}_{sib,g-1} - \boldsymbol{\gamma})$ in group i_g $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g)$. For simplicity, we assume mutual independence among all g vectors in $\mathbf{Y}_{\boldsymbol{\gamma}}^{+}$ so that $\mathbf{Y}_{\boldsymbol{\gamma}}^{+}$ has CDF $F_{\mathbf{Y}}^{(i_1)}(\mathbf{c}_{ind}-\boldsymbol{\gamma}) \times F_{\mathbf{Y}}^{(i_2)}(\mathbf{c}_{sib,1}-\boldsymbol{\gamma}) \times \ldots F_{\mathbf{Y}}^{(i_g)}(\mathbf{c}_{sib,g-1}-\boldsymbol{\gamma}) \equiv F_{\mathbf{Y}^+}^{(l)}(\mathbf{c}_{ind}-\boldsymbol{\gamma}, \mathbf{c}_{sib,1}-\boldsymbol{\gamma}, \ldots, \mathbf{c}_{sib,g-1}-\boldsymbol{\gamma})$ in the l^{th} group ordering $(l = 1, \ldots, g!)$. In the paired case, for example, if $\mathbf{Y}_{\boldsymbol{\gamma}}^{+}$ belongs to the first group ordering, then $\mathbf{Y}_{\boldsymbol{\gamma},ind}$ and $\mathbf{Y}_{\boldsymbol{\gamma},sib,1}$ belong to groups 1 and 2, respectively. Otherwise, $\mathbf{Y}_{\boldsymbol{\gamma},ind}$ belongs to group 2 and $\mathbf{Y}_{\boldsymbol{\gamma},sib,1}$ belongs to group 1. The details on how to construct the tree $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ based on the stacked feature vector $\mathbf{Y}_{\boldsymbol{\gamma}}^{+}$ can be found in Appendix C.2.

In theory, it is possible to construct a tree $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ based on the stacked feature vector $\mathbf{Y}_{\boldsymbol{\gamma}}^{+}$ and use $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ to simultaneously classify all members of a match into one of the $g!$ group orderings, as we show in Appendix C.2. However, $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ does not provide us with direct information that we can use in practice to help us discriminate among the g groups under consideration. The practical problem is that in constructing a tree based on $\mathbf{Y}_{\boldsymbol{\gamma}}^{+}$, the p^{th} feature variable in $\mathbf{Y}_{\boldsymbol{\gamma},ind}$, the p^{th} feature variable in $\mathbf{Y}_{\boldsymbol{\gamma},sib,1}$, ..., and the p^{th} feature variable in $\mathbf{Y}_{\boldsymbol{\gamma},sib,g-1}$ $(p = 1, \ldots, P)$ are treated as g different feature variables, even though they all correspond to the same feature variable. For example, suppose this construction method were applied to the Sweet et al. data. The p^{th} biomarker for each subject in each pair would then be treated as two different biomarkers, even though they both correspond to the same biomarker. As a result, it is possible to construct $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ based on one biomarker that corresponds to a normal control in a given pair and another biomarker that corresponds to a schizophrenia subject in the same pair, which does not make sense from a discriminatory standpoint. Thus, we see in this case that $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ would not yield sensible results that we can use to determine which biomarkers best distinguish between the control and schizophrenia diagnostic groups in a given pair.

Although $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ may be computable and perhaps useful in classification, the fact that

each of the P feature variables of interest is counted g times in the construction of $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ entails that this tree does not produce practical results that we can use to determine the feature variables that best discriminate among the g groups.

4.4.2 Estimation of Unknown Distributions Using Training Data

Due to the key issue that arises in using the stacked approach (in Section 4.4.1.3), we only discuss how to apply the procedures we develop in Sections 4.4.1.1 and 4.4.1.2 using the available training data $\mathbf{y}_{ik}$, the observed feature vector for the member of the k^{th} match belonging to group i ($i = 1, \ldots, g;\ k = 1, \ldots, K$).

4.4.2.1 Tree Construction using Feature Data, adjusting for Effect of Matching

To implement the approach in Section 4.4.1.1 using the available training data, we must first estimate $\boldsymbol{\gamma}$ for the k^{th} match in the training data ($k = 1, \ldots, K$), which we denote as $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,P})'$. Letting $\mathbf{Y}_{ik}$ denote the random feature vector corresponding to the member of the k^{th} match belonging to group i ($i = 1, \ldots, g;\ k = 1, \ldots, K$), we can begin by assuming that for a given match, the mean of $\mathbf{Y}_{ik}$ is given by $E[\mathbf{Y}_{ik}] = \boldsymbol{\lambda}_i + \boldsymbol{\gamma}_k = (\lambda_{1,i} + \gamma_{k,1}, \ldots, \lambda_{P,i} + \gamma_{k,P})'$. Based on the training data, we use LS estimation to estimate $\gamma_{1,p}, \ldots, \gamma_{K,p}, \lambda_{p,1}, \ldots, \lambda_{p,g}$ by minimizing $Q_p = \sum_{i=1}^{g} \sum_{k=1}^{K} (y_{ik,p} - \lambda_{p,i} - \gamma_{k,p})^2$ ($p = 1, \ldots, P$). Using standard arguments, we can easily show that the LS estimates of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\gamma}_k$ are not unique and are given by $\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*) = \bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}}_{..} - \mathbf{c}^*$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*) = \bar{\mathbf{y}}_{.k} + \mathbf{c}^*$, where $\bar{\mathbf{y}}_{i.}$, $\bar{\mathbf{y}}_{.k}$, and $\bar{\mathbf{y}}_{..}$ are defined as in Section 3.5.2.1, $\mathbf{c}^* = -\bar{\mathbf{y}}_{..} + \mathbf{c}$, and $\mathbf{c} \in \mathbb{R}^P$. We see that the estimates of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\gamma}_k$ are the same as the estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\gamma}_k$ in Section 3.5.2.1 for matched LDA.

Once the estimates $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$ are computed for a particular $\mathbf{c}^*$, we can obtain the adjusted training feature measurements $(\hat{\tilde{y}}_{ik,1}, \ldots, \hat{\tilde{y}}_{ik,P})' = \hat{\tilde{\mathbf{y}}}_{ik} = \mathbf{y}_{ik} - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*)$, which can readily be shown to equal $\frac{g-1}{g}\mathbf{D}_{ik,y} - \mathbf{c}^*$, where $\mathbf{D}_{ik,y} = \mathbf{y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l \neq i}}^{g} \mathbf{y}_{lk}$. Based on the adjusted feature data $\hat{\tilde{\mathbf{y}}}_{ik}$, we can estimate the CDF of $\tilde{\mathbf{Y}}_{ind}$ in the i^{th} group ($i = 1, \ldots, g$) as

$$\hat{F}^{(i)}_{\tilde{\mathbf{Y}}_{ind}}(\tilde{\mathbf{c}}) = \hat{P}^{(i)}(\tilde{Y}_{1,ind} \leq \tilde{c}_1, \ldots, \tilde{Y}_{P,ind} \leq \tilde{c}_P) = \frac{\sum_{k=1}^{K} I(\hat{\tilde{y}}_{ik,1} \leq \tilde{c}_1, \ldots, \hat{\tilde{y}}_{ik,P} \leq \tilde{c}_P)}{K}, \quad (4.23)$$

so that we estimate $P^{(i)}(\tilde{\mathbf{Y}}_{ind} \in t)$ as

$$\hat{P}^{(i)}(\tilde{\mathbf{Y}}_{ind} \in t) = \frac{\sum_{k=1}^{K} I(\hat{\tilde{\mathbf{y}}}_{ik} \in t)}{K}. \quad (4.24)$$

Using the priors $\pi_i = 1/g$ and the probability estimates in (4.24), we can apply the standard non-parametric procedure in Section 4.1.3.2 to the adjusted training feature data $\hat{\tilde{\mathbf{y}}}_{ik}$ to construct a tree $T_{\hat{\boldsymbol{\gamma}}}$ that takes into account the effect of group matching and prune $T_{\hat{\boldsymbol{\gamma}}}$ using traditional minimal cost-complexity pruning, which can be implemented using presently available software.

In order to use $T_{\hat{\boldsymbol{\gamma}}}$ to classify each individual in a new match beyond the training data, we must re-estimate $\boldsymbol{\gamma}$ for this match. This is precisely the case where our semi-parametric model in Section 4.3.2 needs to be extended to include parameters that must be re-estimated for each individual in a new match, namely, $\boldsymbol{\gamma}$. If we know the feature measurements for an individual and their $g-1$ siblings in a match, $\mathbf{y}_{ind}$, $\mathbf{y}_{sib,1}$, $\ldots$, $\mathbf{y}_{sib,g-1}$, then we begin by applying our model for the random feature vector $\mathbf{Y}_{ik}$ to the random feature vectors $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ corresponding to a particular match, while retaining the LS estimates $\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*)$ $(i = 1, \ldots, g)$ obtained from the training data. Specifically, for a given match and conditional on the estimates $\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*)$, we assume $\mathbf{Y}_{ind}$ has mean $\hat{\boldsymbol{\lambda}}_{i_1}(\mathbf{c}^*) + \boldsymbol{\gamma}$ in group i_1, $\mathbf{Y}_{sib,1}$ has mean $\hat{\boldsymbol{\lambda}}_{i_2}(\mathbf{c}^*) + \boldsymbol{\gamma}$ in group i_2, $\ldots$, and $\mathbf{Y}_{sib,g-1}$ has mean $\hat{\boldsymbol{\lambda}}_{i_g}(\mathbf{c}^*) + \boldsymbol{\gamma}$ in group i_g $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g)$. We can also examine $\mathbf{Y}_{ind} - \hat{\boldsymbol{\lambda}}_{i_1}(\mathbf{c}^*) \equiv \mathbf{Y}^*_{ind}$, $\mathbf{Y}_{sib,1} - \hat{\boldsymbol{\lambda}}_{i_2}(\mathbf{c}^*) \equiv \mathbf{Y}^*_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1} - \hat{\boldsymbol{\lambda}}_{i_g}(\mathbf{c}^*) \equiv \mathbf{Y}^*_{sib,g-1}$ which all have mean $\boldsymbol{\gamma}$. Next, we use LS estimation to estimate γ_p $(p = 1, \ldots, P)$ by minimizing $Q_p = (y^*_{ind,p} - \gamma_p)^2 + \sum_{m=1}^{g-1}(y^*_{sib,m,p} - \gamma_p)^2$ so that the LS estimate of $\boldsymbol{\gamma}$ is given by

$$\hat{\boldsymbol{\gamma}}(\mathbf{c}^*) = \frac{1}{g}(\mathbf{y}^*_{ind} + \sum_{m=1}^{g-1} \mathbf{y}^*_{sib,m}) = \frac{1}{g}\left[(\mathbf{y}_{ind} + \sum_{m=1}^{g-1} \mathbf{y}_{sib,m}) - \sum_{i=1}^{g} \hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*)\right], \tag{4.25}$$

which can be shown to equal $\frac{1}{g}(\mathbf{y}_{ind} + \sum_{m=1}^{g-1} \mathbf{y}_{sib,m}) + \mathbf{c}^*$. No matter which groups $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ belong to, the LS criteria Q_p remain invariant and, thus, $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$ remains the same. From (4.25), we see that our estimate of $\boldsymbol{\gamma}$ for a new match is identical to the estimate of $\boldsymbol{\gamma}$ for each match in the training data, namely, $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*) = \bar{\mathbf{y}}_{\cdot k} + \mathbf{c}^*$. The adjusted tree $T_{\hat{\boldsymbol{\gamma}}}$ can then be used to classify each individual in any given match based on their adjusted value $\tilde{\mathbf{y}}_{ind} = \mathbf{y}_{ind} - \hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$, which is equal to $\frac{g-1}{g}\mathbf{y}_{\text{diff}} - \mathbf{c}^*$, where $\mathbf{y}_{\text{diff}} = \mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1} \mathbf{y}_{sib,m}$.

On the other hand, if we only know the feature measurement $\mathbf{y}_{ind}$ for an individual in a new match, then we cannot estimate $\boldsymbol{\gamma}$ for this individual in the manner we just described. Without an estimate of $\boldsymbol{\gamma}$, it's clear that we cannot use $T_{\hat{\boldsymbol{\gamma}}}$ to classify this individual. As we

suggested previously, in studies where subjects are matched across the g groups of interest, such as post-mortem tissue studies, it makes no sense to try to classify a single observation of a match without knowing the remaining $g-1$ observations in that match.

We now discuss several properties of our adjusted tree $T_{\hat{\boldsymbol{\gamma}}}$, which follow from the linear invariance property (Proposition 4.3.1.1). First, the set of splitting variables chosen for $T_{\hat{\boldsymbol{\gamma}}}$ does not change depending on $\mathbf{c}^*$. Also, suppose we consider any of the P feature variables in $\tilde{\mathbf{Y}}_{ind}$, namely, $\tilde{Y}_{\nu,ind}$ ($\nu \in (1,2,\ldots,P)$), and the trees $T_{\hat{\boldsymbol{\gamma}},a}$ and $T_{\hat{\boldsymbol{\gamma}},b}$ constructed using the estimates $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}_a^*)$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}_b^*)$, respectively, where $\mathbf{c}_a^* = -\bar{\mathbf{y}}_{\cdot\cdot} + \mathbf{c}_a$, $\mathbf{c}_b^* = -\bar{\mathbf{y}}_{\cdot\cdot} + \mathbf{c}_b$, and $\mathbf{c}_a \neq \mathbf{c}_b$. We then have that the split $\tilde{Y}_{\nu,ind,a} \leq \tilde{c}_{\nu,a}$ in $T_{\hat{\boldsymbol{\gamma}},a}$ is equivalent to the split $\tilde{Y}_{\nu,ind,b} \leq \tilde{c}_{\nu,b}$ in $T_{\hat{\boldsymbol{\gamma}},b}$, where $\tilde{Y}_{\nu,ind,a} = Y_{\nu,ind} - \hat{\gamma}_{\nu,a}$, $\tilde{Y}_{\nu,ind,b} = Y_{\nu,ind} - \hat{\gamma}_{\nu,b}$, $\hat{\gamma}_{\nu,a}$ and $\hat{\gamma}_{\nu,b}$ are the ν^{th} elements of $\hat{\boldsymbol{\gamma}}(\mathbf{c}_a^*)$ and $\hat{\boldsymbol{\gamma}}(\mathbf{c}_b^*)$, and $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*)$ is defined as in (4.25). These two splits are equivalent due to the fact that $\tilde{Y}_{\nu,ind,b} = \tilde{Y}_{\nu,ind,a} + \hat{\gamma}_{\nu,a} - \hat{\gamma}_{\nu,b}$ and $\tilde{c}_{\nu,b} = \tilde{c}_{\nu,a} + \hat{\gamma}_{\nu,a} - \hat{\gamma}_{\nu,b}$. Since the splits $\tilde{Y}_{\nu,ind,a} \leq \tilde{c}_{\nu,a}$ and $\tilde{Y}_{\nu,ind,b} \leq \tilde{c}_{\nu,b}$ are equivalent, so are the splits $Y_{\nu,ind} \leq \tilde{c}_{\nu,a} + \hat{\gamma}_{\nu,a}$ and $Y_{\nu,ind} \leq \tilde{c}_{\nu,b} + \hat{\gamma}_{\nu,b}$, which implies that $\tilde{c}_{\nu,a} + \hat{\gamma}_{\nu,a} \equiv \tilde{c}_{\nu,b} + \hat{\gamma}_{\nu,b}$. In other words, the following can be said of $T_{\hat{\boldsymbol{\gamma}}}$: (1) regardless of the value of $\mathbf{c}^*$, $T_{\hat{\boldsymbol{\gamma}}}$ yields the same classification results; (2) the set of optimal cutpoints for $T_{\hat{\boldsymbol{\gamma}}}$ changes depending on $\mathbf{c}^*$ when $T_{\hat{\boldsymbol{\gamma}}}$ is expressed in terms of the adjusted feature data $\tilde{\mathbf{Y}}_{ind}$, but not when $T_{\hat{\boldsymbol{\gamma}}}$ is expressed in terms of the original feature data $\mathbf{Y}_{ind}$.

4.4.2.2 Tree Construction using Differenced Feature Vector

Alternately, we can apply the differencing approach in Section 4.4.1.2 using training data. In this case, we begin by estimating the CDFs $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\centerdot)$ in the i^{th} population ($i = 1,\ldots,g$) from the training feature differences $(D_{ik,y,1},\ldots,D_{ik,y,P})' = \mathbf{D}_{ik,y} = \mathbf{y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g} \mathbf{y}_{lk}$, where $\mathbf{D}_{ik,y}$ is computed for the i^{th} group member of the k^{th} match ($i = 1,\ldots,g$; $k = 1,\ldots,K$). Based on our model for $\mathbf{Y}_{\text{diff}}$ in Section 4.4.1.2, the training feature difference $\mathbf{D}_{ik,y}$ belongs to the i^{th} population. We can then estimate $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\centerdot)$ by computing the empirical CDF

$$\hat{F}^{(i)}_{\mathbf{Y}_{\text{diff}}}(\mathbf{c}) = \hat{P}^{(i)}(Y_{1,\text{diff}} \leq c_1,\ldots,Y_{P,\text{diff}} \leq c_P) = \frac{1}{K}\sum_{k=1}^{K} I(D_{ik,y,1} \leq c_1,\ldots,D_{ik,y,P} \leq c_P), \tag{4.26}$$

from which $P^{(i)}(\mathbf{Y}_{\text{diff}} \in t)$ can be estimated as

$$\hat{P}^{(i)}(\mathbf{Y}_{\text{diff}} \in t) = \frac{\sum_{k=1}^{K} I(\mathbf{D}_{ik,y} \in t)}{K}. \tag{4.27}$$

Based on the probability estimates in (4.27) and our assumption of equal priors, we can implement the standard non-parametric construction approach in Section 4.1.3.2 on the training feature differences $\mathbf{D}_{ik,y}$ to construct a tree T_{diff} that adjusts for the effect of group matching and use minimal cost-complexity pruning to prune T_{diff} accordingly, which we can carry out using available software.

In comparing the trees $T_{\hat{\boldsymbol{\gamma}}}$ and T_{diff} constructed using $\hat{\tilde{\mathbf{y}}}_{ik}$ and $\mathbf{D}_{ik,y}$, respectively, we point out the following facts, which follow from the linear invariance property (Proposition 4.3.1.1). First, since $\hat{\tilde{\mathbf{y}}}_{ik} = \frac{g-1}{g}\mathbf{D}_{ik,y} - \mathbf{c}^*$, as we showed in Section 4.4.2.1, we have that $\hat{\tilde{\mathbf{y}}}_{ik}$ is an increasing linear function of $\mathbf{D}_{ik,y}$ and, thus, the same set of feature variables is chosen for $T_{\hat{\boldsymbol{\gamma}}}$ and T_{diff}. Also, if $Y_{\nu,\text{diff}}$ and $\tilde{Y}_{\nu,ind}$ correspond to any of the P feature variables in $\mathbf{Y}_{\text{diff}}$ and $\tilde{\mathbf{Y}}_{ind}$ ($\nu \in (1, 2, \ldots, P)$), respectively, then the split $Y_{\nu,\text{diff}} \leq c_\nu$ in T_{diff} is equivalent to the split $\tilde{Y}_{\nu,ind} \leq \frac{g-1}{g}c_\nu - c_\nu^*$ in $T_{\hat{\boldsymbol{\gamma}}}$, where c_ν^* is the ν^{th} element of $\mathbf{c}^*$ and $\mathbf{c}^*$ is defined as in Section 4.4.2.1. As a result, $T_{\hat{\boldsymbol{\gamma}}}$ and T_{diff} produce the same classification results and, more importantly, identify the same set of feature variables that best discriminate among the g groups under consideration, once the effect of group matching on the feature data is accounted for.

4.5 MATCHED CLASSIFICATION TREES WITH COVARIATES

4.5.1 Known Distributions

In Section 4.5, we extend our construction methodology for matched classification trees to also account for the effects of additional covariates on the feature data. Although Section 4.5 is included for completeness, the details are very similar to Section 4.4 and, thus, the reader can safely skip this section.

Retaining our assumptions from Section 4.4 of equal priors and equal misclassification costs, we begin with an extension of our population-based procedures in Sections 4.4.1.1 and 4.4.1.2 to also account for covariate effects. Due to the fact that the construction approach in Section 4.4.1.3 based on the stacked feature data $\mathbf{Y}_{\boldsymbol{\gamma}}^+$ does not produce results that are useful for our purposes, mainly those pertaining to group discrimination, we do not provide

an extension of this approach.

We note that we could choose to ignore the effect of matching and apply our semi-parametric tree construction procedure to instead account for the effects of the variables on which individuals are matched, e.g., age at death, gender, and PMI in post-mortem tissue studies. When we ignore matching, however, we only consider the feature data for one randomly selected individual in the population, rather than the feature data for g individuals in a particular match. In this case, the feature vector $\mathbf{Y}$ for this randomly selected individual may not necessarily have an equal probability of belonging to any of the g groups of interest, as is the case if we know that $\mathbf{Y}$ corresponds to an individual in a match. Thus, in the unmatched case, an assumption of equal priors across the g groups may not be appropriate in certain contexts. For example, we may find it more appropriate to use the population proportion of normal controls and that of individuals with schizophrenia.

4.5.1.1 Tree Construction using Feature Vector, adjusting for Matching and Covariate Effects Let $\mathbf{X}_{ind}$ denote the random covariate vector an individual in a match. For known $\boldsymbol{\gamma}$ and given $\mathbf{X}_{ind} = \mathbf{x}_{ind}$, let $\mathbf{Y}_{\boldsymbol{\gamma},\mathbf{x}_{ind}}$ denote the random feature vector for any individual belonging to a match, conditional on $\mathbf{x}_{ind}$. We let $\mathbf{Y}_{\boldsymbol{\gamma},\mathbf{x}_{ind}}$ have known conditional CDF $F^{(i)}_{\mathbf{Y}_{ind}}(\mathbf{c}-\boldsymbol{\gamma}-\boldsymbol{\beta}\mathbf{x}_{ind})$ in the i^{th} group $(i = 1, \ldots, g)$, where $\mathbf{Y}_{ind}$ is defined as in Section 4.4.1.1 and $\boldsymbol{\beta}$ is known and is defined as in Section 3.4.1.1. In other words, $\mathbf{Y}_{\boldsymbol{\gamma},\mathbf{x}_{ind}} \stackrel{d}{=} \mathbf{Y}_{ind} + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{ind}$. Equivalently, we may examine the random feature vector that has been adjusted for both matching and covariate effects, namely, $\tilde{\mathbf{Y}}_{ind(x)} = \mathbf{Y}_{\boldsymbol{\gamma},\mathbf{x}_{ind}} - \boldsymbol{\gamma} - \boldsymbol{\beta}\mathbf{x}_{ind}$ with known CDF $F^{(i)}_{\tilde{\mathbf{Y}}_{ind(x)}}(\centerdot)$ in the i^{th} group, where $F^{(i)}_{\tilde{\mathbf{Y}}_{ind(x)}}(\centerdot) \equiv F^{(i)}_{\mathbf{Y}_{ind}}(\centerdot)$. Although we only consider a linear function of the covariate data in our discussion, we can easily generalize to the case where conditional on $\mathbf{x}_{ind}$, $\mathbf{Y}_{\boldsymbol{\gamma},\mathbf{x}_{ind}}$ has CDF $F^{(i)}_{\mathbf{Y}_{ind}}(\mathbf{c}-\boldsymbol{\gamma}-\boldsymbol{\xi}(\mathbf{x}_{ind};\boldsymbol{\Theta}))$, where $\boldsymbol{\xi}(\mathbf{x};\boldsymbol{\Theta})$ is defined as in Section 4.3.2.

Based on the distribution of $\tilde{\mathbf{Y}}_{ind(x)}$, and our assumption of equal priors and misclassification costs, we can use the traditional population-based approach in Section 4.1.2.1 to construct a tree $T'^{\mathrm{adj}(\boldsymbol{\gamma},\mathbf{x})}$ that adjusts for matching and covariate effects by replacing the probabilities $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ used to construct T' with the probabilities $P^{(i)}(\tilde{\mathbf{Y}}_{ind(x)} \in t)$, $P^{(i)}(\tilde{\mathbf{Y}}_{ind(x)} \in t_L)$, and $P^{(i)}(\tilde{\mathbf{Y}}_{ind(x)} \in t_R)$ obtained from $F^{(i)}_{\tilde{\mathbf{Y}}_{ind(x)}}(\centerdot)$. We can then use the adjusted tree $T'^{\mathrm{adj}(\boldsymbol{\gamma},\mathbf{x})}$ to classify any individual in a match based on

their adjusted feature measurement $\tilde{\mathbf{y}}_{ind(x)}$. In addition, we can use this tree to determine the feature variables, once adjusted for matching and covariate effects, and corresponding splits that best differentiate among the g groups of interest. For example, in the context of post-mortem tissue studies, this tree can be used to identify the biomarkers and corresponding splits that best discriminate between the control and schizophrenia diagnostic groups, once we account for the effects of both subject pairing and additional covariates, such as brain tissue storage time, on the biomarker data.

4.5.1.2 Tree Construction using Covariate Adjusted Differenced Feature Vector

Another approach we can take to account for both matching and covariate effects is to apply our semi-parametric tree construction procedure to the differenced random feature vector $\mathbf{Y}_{\text{diff}} \equiv \mathbf{Y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{Y}_{sib,m}$.

To elaborate, we first let $\mathbf{X}_{ind}, \mathbf{X}_{sib,1}, \ldots, \mathbf{X}_{sib,g-1}$ denote the random covariate vectors for any individual and their $g-1$ siblings in a match, and let $\mathbf{X}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{X}_{sib,m} \equiv \mathbf{X}_{\text{diff}}$ denote the differenced random covariate vector for any individual in a match. Given $\mathbf{X}_{\text{diff}} = \mathbf{x}_{\text{diff}}$, we let $\mathbf{Y}_{\text{diff(x)}}$ denote the differenced random feature vector conditional on $\mathbf{x}_{\text{diff}}$ with known CDF $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\mathbf{c} - \boldsymbol{\beta}\mathbf{x}_{\text{diff}})$ in the i^{th} population $(i = 1, \ldots, g)$, where $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\mathbf{c})$ are defined as in (4.21) and $\boldsymbol{\beta}$ is known. As a result, $\mathbf{Y}_{\text{diff(x)}} \overset{d}{=} \mathbf{Y}_{\text{diff}} + \boldsymbol{\beta}\mathbf{x}_{\text{diff}}$. We note that if $\mathbf{Y}_{\text{diff(x)}}$ belongs to the i^{th} population, then $\mathbf{Y}_{ind}$ belongs to group i.

We may also consider the covariate adjusted differenced feature vector $\tilde{\mathbf{Y}}_{\text{diff(x)}} = \mathbf{Y}_{\text{diff}(\mathbf{x})} - \boldsymbol{\beta}\mathbf{x}_{\text{diff}}$ with known CDF $F^{(i)}_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}(\centerdot)$ in the i^{th} population, where $F^{(i)}_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}(\centerdot) \equiv F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\centerdot)$. Since the CDFs $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\centerdot)$ and $F^{(i)}_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}(\centerdot)$ are the same, the conditional probabilities $P^{(i)}_{\mathbf{x}_{\text{diff}}}(\mathbf{Y}_{\text{diff}(\mathbf{x})} \in t) = P_{\mathbf{x}_{\text{diff}}}(\mathbf{Y}_{\text{diff}(\mathbf{x})} \in t | \mathbf{Y}_{\text{diff}(\mathbf{x})} \in \text{population } i)$ obtained from $F^{(i)}_{\mathbf{Y}_{\text{diff}}}(\mathbf{c} - \boldsymbol{\beta}\mathbf{x}_{\text{diff}})$ can also be expressed as $P^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t) = P(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t | \tilde{\mathbf{Y}}_{\text{diff(x)}} \in \text{population } i)$ obtained from $F^{(i)}_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}(\centerdot)$.

Retaining our assumption from Section 4.4.1.2 of equal priors across the g populations, we can construct our adjusted tree $T'^{(\text{diff}(\mathbf{x}))}$ using the traditional procedure in Section 4.1.2.1 by replacing $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ used to construct T' with $P^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t)$, $P^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t_L)$, and $P^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t_R)$. The following rule is then used to assign each terminal node t of $T'^{(\text{diff}(\mathbf{x}))}$ to the i^{th} population of $\tilde{\mathbf{Y}}_{\text{diff(x)}}$:

$$R^{\text{diff}}_{i(x)} : \left\{ t : P^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t) > P^{(j)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t) \right\}, \quad j = 1, \ldots, g;\ j \neq i. \tag{4.28}$$

For each individual in a match, we compute the covariate adjusted difference $\tilde{\mathbf{y}}_{\text{diff(x)}} = (\mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{y}_{sib,m}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{x}_{sib,m})$. If this difference falls into a terminal node of

$T'^{(\text{diff}(\mathbf{x}))}$ that has been assigned to the i^{th} population based on the rule in (4.28), then we classify that individual into the i^{th} group ($i = 1, \ldots, g$).

Using the adjusted tree $T'^{(\text{diff}(\mathbf{x}))}$, we can identify among the feature variables of interest those that best discriminate among the g groups, once we account for group matching and covariate effects on the feature data. As we did in Section 4.4.1.2 when we only accounted for the effect of matching, we briefly discuss what we can determine from $T'^{(\text{diff}(\mathbf{x}))}$ when we're dealing with matched pairs. First, we point out that the covariate adjusted difference $\tilde{\mathbf{Y}}_{\text{diff(x)}}$ can also be written as $(\mathbf{Y}_{ind} - \boldsymbol{\beta}\mathbf{x}_{ind}) - (\mathbf{Y}_{sib} - \boldsymbol{\beta}\mathbf{x}_{sib})$, which is the difference between the covariate adjusted random feature vector for an individual in a pair and that of their sibling. When we view $\tilde{\mathbf{Y}}_{\text{diff(x)}}$ in this manner, we have that in the case of matched pairs, the adjusted tree $T'^{(\text{diff}(\mathbf{x}))}$ helps us identify the feature variables that best distinguish an individual belonging to group 1 from that belonging to group 2 in any given pair, once these feature variables have been suitably adjusted for covariate effects. Also, as was the case for $T'^{(\text{diff})}$, the cutpoints of $T'^{(\text{diff}(\mathbf{x}))}$ can be used to determine whether large or small values of each covariate adjusted splitting variable in $T'^{(\text{diff}(\mathbf{x}))}$ for an individual in a pair, relative to the values of the same adjusted splitting variable for the individual's sibling in that pair, are associated with group 1 compared with group 2.

4.5.2 Estimation of Unknown Distributions Using Training Data

With available training data consisting of $(\mathbf{y}_{ik}, \mathbf{x}_{ik})$, the observed feature and covariate vectors for the member of the k^{th} match belonging to group i ($i = 1, \ldots, g$; $k = 1, \ldots, K$), we now discuss how to estimate the adjusted CDFs $F^{(i)}_{\tilde{\mathbf{Y}}_{ind(x)}}(\centerdot)$ and $F^{(i)}_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}(\centerdot)$ in Sections 4.5.1.1 and 4.5.1.2, respectively.

4.5.2.1 Tree Construction using Feature Data, adjusting for Matching and Covariate Effects

We begin by first estimating the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for each match in the training data. One approach we can take is to assume that the conditional mean of the random feature vector $\mathbf{Y}_{ik}$ is given by $E[\mathbf{Y}_{ik}|\mathbf{x}_{ik}] = \boldsymbol{\lambda}_i + \boldsymbol{\gamma}_k + \boldsymbol{\beta}\mathbf{x}_{ik} =$
$(\lambda_{1,i} + \gamma_{k,1} + \boldsymbol{\beta}_1\mathbf{x}_{ik}, \ldots, \lambda_{P,i} + \gamma_{k,P} + \boldsymbol{\beta}_P\mathbf{x}_{ik})'$. Based on the training data, we use LS estimation to estimate $\boldsymbol{\beta}_p$, $\gamma_{1,p}$, $\ldots$, $\gamma_{K,p}$, $\lambda_{p,1}$, $\ldots$, $\lambda_{p,g}$ by minimizing $Q_p = \sum_{i=1}^{g}\sum_{k=1}^{K}(y_{ik,p} -$

$\lambda_{p,i} - \gamma_{k,p} - \boldsymbol{\beta}_p \mathbf{x}_{ik})^2$ $(p = 1, \ldots, P)$. Assuming that the design matrix for our model satisfies suitable conditions so that the LS estimate $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_P \end{bmatrix}$ is unique, the LS estimates of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\gamma}_k$ are given by $\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*_{\mathbf{x}}) = \bar{\mathbf{y}}_{i\cdot} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{i\cdot} - (\bar{\mathbf{y}}_{\cdot\cdot} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\cdot\cdot}) - \mathbf{c}^*_{\mathbf{x}}$ and $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) = \bar{\mathbf{y}}_{\cdot k} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\cdot k} + \mathbf{c}^*_{\mathbf{x}}$, which are the same as the estimates of $\boldsymbol{\mu}_i$ and $\boldsymbol{\gamma}_k$ in Section 3.6.2.1 for matched LDA with covariates.

For a specific $\mathbf{c}^*_{\mathbf{x}}$, we can compute the training feature data that have been adjusted for matching and covariate effects $(\hat{\tilde{y}}_{ik(x),1}, \ldots, \hat{\tilde{y}}_{ik(x),P})' = \hat{\tilde{\mathbf{y}}}_{ik(x)} = \mathbf{y}_{ik} - \hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ik}$, which can easily be shown to equal $\frac{g-1}{g}\hat{\mathbf{D}}^{adj}_{ik,y} - \mathbf{c}^*_{\mathbf{x}}$, where $\hat{\mathbf{D}}^{adj}_{ik,y} = (\mathbf{y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g} \mathbf{y}_{lk}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g} \mathbf{x}_{lk})$ $(i = 1, \ldots, g)$. From the adjusted training feature data, we can estimate the CDF of $\tilde{\mathbf{Y}}_{ind(x)}$ in the i^{th} group as

$$\hat{F}^{(i)}_{\tilde{\mathbf{Y}}_{ind(x)}}(\tilde{\mathbf{c}}) = \hat{P}^{(i)}(\tilde{Y}_{1,ind(x)} \leq \tilde{c}_1, \ldots, \tilde{Y}_{P,ind(x)} \leq \tilde{c}_P) = \frac{\sum_{k=1}^{K} I(\hat{\tilde{y}}_{ik(x),1} \leq \tilde{c}_1, \ldots, \hat{\tilde{y}}_{ik(x),P} \leq \tilde{c}_P)}{K} \tag{4.29}$$

and the probabilities $P^{(i)}(\tilde{\mathbf{Y}}_{ind(x)} \in t)$ as

$$\hat{P}^{(i)}(\tilde{\mathbf{Y}}_{ind(x)} \in t) = \frac{\sum_{k=1}^{K} I(\hat{\tilde{\mathbf{y}}}_{ik(x)} \in t)}{K}. \tag{4.30}$$

Based on the priors $\pi_i = 1/g$ and the probability estimates in (4.30), we can implement the standard non-parametric approach in Section 4.1.3.2 on the adjusted training feature measurements $\hat{\tilde{\mathbf{y}}}_{ik(x)}$ to construct a tree $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ that adjusts for the effects of both group matching and additional covariates on the feature data, and then prune this adjusted tree using minimal cost-complexity pruning. To construct $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ in this manner, we can use several software packages, e.g., R or Salford Systems CART®.

To use the tree $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ to classify each member of a new match, we must have an estimate of $\boldsymbol{\gamma}$ for this match, which we can obtain if we know the feature and covariate measurements for an individual and their $g-1$ siblings in that match, $(\mathbf{y}_{ind}, \mathbf{x}_{ind})$, $(\mathbf{y}_{sib,1}, \mathbf{x}_{sib,1})$, $\ldots$, $(\mathbf{y}_{sib,g-1}, \mathbf{x}_{sib,g-1})$. Based on the LS estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*_{\mathbf{x}})$ from the training data, we apply the conditional model for the random feature vector $\mathbf{Y}_{ik}$ to the random feature vectors $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}, \ldots, \mathbf{Y}_{sib,g-1}$ for a new match. To clarify, for any given match and given $\mathbf{x}_{ind}$, $\mathbf{x}_{sib,1}, \ldots, \mathbf{x}_{sib,g-1}$, $\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*)$, and $\hat{\boldsymbol{\beta}}$, we assume $\mathbf{Y}_{ind}$ has conditional mean $\hat{\boldsymbol{\lambda}}_{i_1}(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{ind}$ in group i_1, $\mathbf{Y}_{sib,1}$ has conditional mean $\hat{\boldsymbol{\lambda}}_{i_2}(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,1}$ in group i_2, $\ldots$, and $\mathbf{Y}_{sib,g-1}$ has conditional mean $\hat{\boldsymbol{\lambda}}_{i_g}(\mathbf{c}^*_{\mathbf{x}}) + \boldsymbol{\gamma} + \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,g-1}$ in group i_g $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g)$.

Alternately, we can consider $\mathbf{Y}_{ind} - \hat{\boldsymbol{\lambda}}_{i_1}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ind} \equiv \mathbf{Y}^{*(x)}_{ind}$, $\mathbf{Y}_{sib,1} - \hat{\boldsymbol{\lambda}}_{i_2}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,1} \equiv \mathbf{Y}^{*(x)}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1} - \hat{\boldsymbol{\lambda}}_{i_g}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{sib,g-1} \equiv \mathbf{Y}^{*(x)}_{sib,g-1}$ which all have mean $\boldsymbol{\gamma}$. We then estimate γ_p by minimizing the LS criteria $Q_p = (y^{*(x)}_{ind,p} - \gamma_p)^2 + \sum_{m=1}^{g-1}(y^{*(x)}_{sib,m,p} - \gamma_p)^2$ $(p = 1, \ldots, P)$ so that the LS estimate of $\boldsymbol{\gamma}$ is given by

$$\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}}) = \frac{1}{g}(\mathbf{y}^{*(x)}_{ind} + \sum_{m=1}^{g-1}\mathbf{y}^{*(x)}_{sib,m}) = \frac{1}{g}\left[(\mathbf{y}_{ind} + \sum_{m=1}^{g-1}\mathbf{y}_{sib,m}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} + \sum_{m=1}^{g-1}\mathbf{x}_{sib,m}) - \sum_{i=1}^{g}\hat{\boldsymbol{\lambda}}_i(\mathbf{c}^*_{\mathbf{x}})\right], \tag{4.31}$$

which can be shown to equal $\frac{1}{g}[(\mathbf{y}_{ind} + \sum_{m=1}^{g-1}\mathbf{y}_{sib,m}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} + \sum_{m=1}^{g-1}\mathbf{x}_{sib,m})] + \mathbf{c}^*_{\mathbf{x}}$. As in Section 4.4.2.1, we have that no matter which groups $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, $\ldots$, $\mathbf{Y}_{sib,g-1}$ belong to, the LS criteria Q_p remain invariant and $\hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}})$ remains the same. Also, we see that our estimate of $\boldsymbol{\gamma}$ in (4.31) for a new match is the same as the estimate of $\boldsymbol{\gamma}$ for each training match, i.e., $\hat{\boldsymbol{\gamma}}_k(\mathbf{c}^*_{\mathbf{x}}) = \bar{\mathbf{y}}_{\cdot k} - \hat{\boldsymbol{\beta}}\bar{\mathbf{x}}_{\cdot k} + \mathbf{c}^*_{\mathbf{x}}$. We can then use our adjusted tree $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ to classify each individual in a given match based on their adjusted value $\tilde{\mathbf{y}}_{ind(x)} = \mathbf{y}_{ind} - \hat{\boldsymbol{\gamma}}(\mathbf{c}^*_{\mathbf{x}}) - \hat{\boldsymbol{\beta}}\mathbf{x}_{ind}$, which can be shown to equal $\frac{g-1}{g}\tilde{\mathbf{y}}_{\text{diff(x)}} - \mathbf{c}^*_{\mathbf{x}}$, where $\tilde{\mathbf{y}}_{\text{diff(x)}} = (\mathbf{y}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{y}_{sib,m}) - \hat{\boldsymbol{\beta}}(\mathbf{x}_{ind} - \frac{1}{g-1}\sum_{m=1}^{g-1}\mathbf{x}_{sib,m})$.

Based on the linear invariance property (Proposition 4.3.1.1), we note the following facts regarding the adjusted tree $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$. As was the case for the tree $T_{\hat{\boldsymbol{\gamma}}}$ in Section 4.4.2.1, the set of splitting variables chosen for $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ remains the same regardless of the value of $\mathbf{c}^*_{\mathbf{x}}$. In addition, using an argument similar to the one made in Section 4.4.2.1 for $T_{\hat{\boldsymbol{\gamma}}}$, we have that $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ yields the same classification results regardless of $\mathbf{c}^*_{\mathbf{x}}$ and that the set of optimal cutpoints for $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ depends on $\mathbf{c}^*_{\mathbf{x}}$ when $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ is expressed in terms of the adjusted feature variables $\tilde{Y}_{1,ind(x)}, \ldots, \tilde{Y}_{P,ind(x)}$, but not when $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ is expressed in terms of the original feature variables $Y_{1,ind(x)}, \ldots, Y_{P,ind(x)}$.

4.5.2.2 Tree Construction using Covariate Adjusted Differenced Feature Vector

Another estimation approach we can take is to implement the differencing approach we develop in Section 4.5.1.2 using the available training data. First, we let $\mathbf{D}_{ik,Y} \equiv \mathbf{Y}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g}\mathbf{Y}_{lk}$ denote the random differenced feature vector for the member of the k^{th} match belonging to the i^{th} group $(i = 1, \ldots, g;\ k = 1, \ldots, K)$, such that the difference $\mathbf{D}_{ik,Y}$ belongs to the i^{th} population. Using our conditional model for the differenced random feature vector $\mathbf{Y}_{\text{diff}}$ in Section 4.5.1.2, we assume that the conditional mean of $\mathbf{D}_{ik,Y}$ is given

by $E[\mathbf{D}_{ik,Y}|\mathbf{D}_{ik,x}] = \boldsymbol{\lambda}_i^{\text{diff}} + \boldsymbol{\beta}\mathbf{D}_{ik,x} = \left(\lambda_{1,i}^{\text{diff}} + \boldsymbol{\beta}_1\mathbf{D}_{ik,x}, \ldots, \lambda_{P,i}^{\text{diff}} + \boldsymbol{\beta}_P\mathbf{D}_{ik,x}\right)'$, where $\mathbf{D}_{ik,x} = \mathbf{x}_{ik} - \frac{1}{g-1}\sum_{\substack{l=1 \\ l\neq i}}^{g}\mathbf{x}_{lk}$. Based on the differenced training data $(\mathbf{D}_{ik,y}, \mathbf{D}_{ik,x})$, we use LS estimation to estimate $\boldsymbol{\beta}_p$, $\lambda_{p,1}^{\text{diff}}, \ldots, \lambda_{p,g}^{\text{diff}}$ by minimizing $Q_p = \sum_{i=1}^{g}\sum_{k=1}^{K}(D_{ik,y,p} - \lambda_{p,i}^{\text{diff}} - \boldsymbol{\beta}_p\mathbf{D}_{ik,x})^2$, where $D_{ik,y,p}$ is the p^{th} element of $\mathbf{D}_{ik,y}$ $(p = 1, \ldots, P)$. We assume that the design matrix for our model satisfies suitable conditions so that the LS estimate $\hat{\boldsymbol{\beta}}$ is unique.

Once we obtain the estimate $\hat{\boldsymbol{\beta}}$, our next step is to compute the covariate adjusted training feature differences $(\hat{D}_{ik,y,1}^{adj}, \ldots, \hat{D}_{1k,y,P}^{adj})' = \hat{\mathbf{D}}_{ik,y}^{adj} = \mathbf{D}_{ik,y} - \hat{\boldsymbol{\beta}}\mathbf{D}_{ik,x}$ $(i = 1, \ldots, g)$. Based on these adjusted differences, we can estimate the CDF $F_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}^{(i)}(\centerdot)$ in the i^{th} population by computing the empirical CDF

$$\hat{F}_{\tilde{\mathbf{Y}}_{\text{diff(x)}}}^{(i)}(\tilde{\mathbf{c}}) = \hat{P}^{(i)}(\tilde{Y}_{1,\text{diff(x)}} \leq \tilde{c}_1, \ldots, \tilde{Y}_{P,\text{diff(x)}} \leq \tilde{c}_P) = \frac{1}{K}\sum_{k=1}^{K} I(\hat{D}_{ik,y,1}^{adj} \leq \tilde{c}_1, \ldots, \hat{D}_{ik,y,P}^{adj} \leq \tilde{c}_P), \tag{4.32}$$

from which we can estimate $P^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t)$ as

$$\hat{P}^{(i)}(\tilde{\mathbf{Y}}_{\text{diff(x)}} \in t) = \frac{\sum_{k=1}^{K} I(\hat{\mathbf{D}}_{ik,y}^{adj} \in t)}{K}. \tag{4.33}$$

Using the probability estimates in (4.33) and our assumption of equal priors, we can implement the standard non-parametric approach in Section 4.1.3.2 on the covariate adjusted training feature differences $\hat{\mathbf{D}}_{ik,y}^{adj}$ to construct a tree $T_{\text{diff}(\mathbf{x})}$ that accounts for both matching and covariate effects, and use minimal cost-complexity pruning to prune this tree.

When we compare the trees $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ and $T_{\text{diff}(\mathbf{x})}$ constructed using $\hat{\tilde{\mathbf{y}}}_{ik(x)}$ and $\hat{\mathbf{D}}_{ik,y}^{adj}$, respectively, we can state the following based on the linear invariance property (Proposition 4.3.1.1). First, due to the fact that $\hat{\tilde{\mathbf{y}}}_{ik(x)} = \frac{g-1}{g}\hat{\mathbf{D}}_{ik,y}^{adj} - \mathbf{c}_{\mathbf{x}}^*$, as we showed in Section 4.5.2.1, $\hat{\tilde{\mathbf{y}}}_{ik(x)}$ is an increasing linear function of $\hat{\mathbf{D}}_{ik,y}^{adj}$ and, thus, the same set of feature variables is chosen for $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ and $T_{\text{diff}(\mathbf{x})}$. In addition, suppose $\tilde{Y}_{\nu,\text{diff(x)}}$ and $\tilde{Y}_{\nu,ind(x)}$ denote the ν^{th} feature variables in $\tilde{\mathbf{Y}}_{\text{diff(x)}}$ and $\tilde{\mathbf{Y}}_{ind(x)}$ $(\nu \in (1, 2, \ldots, P))$, respectively. We then have that the split $\tilde{Y}_{\nu,\text{diff(x)}} \leq \tilde{c}_\nu$ in $T_{\text{diff}(\mathbf{x})}$ is equivalent to the split $\tilde{Y}_{\nu,ind(x)} \leq \frac{g-1}{g}\tilde{c}_\nu - c_{\nu,\mathbf{x}}^*$ in $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$, where $c_{\nu,\mathbf{x}}^*$ is the ν^{th} element of $\mathbf{c}_{\mathbf{x}}^*$ and $\mathbf{c}_{\mathbf{x}}^*$ is defined as in Section 4.5.2.1. Therefore, the adjusted trees $T_{\hat{\boldsymbol{\gamma}},\mathbf{x}}$ and $T_{\text{diff}(\mathbf{x})}$ both yield identical classification results and identify the same set of feature variables that best distinguishes among the g groups of interest, once the effects of both group matching and covariates on the feature data have been taken into account.

5.0 APPLICATIONS TO POST-MORTEM TISSUE DATA

5.1 SWEET DATA

5.1.1 Description of Dataset

We first explore our adjustment methodology for LDA and classification trees of our motivating post-mortem tissue data set obtained from the four post-mortem tissue studies conducted by Sweet et al. [33][34][35][36]. In total, six biomarkers, listed in Table 5.1, were measured by Dr. Sweet and his collaborators in post-mortem human brain tissue taken from the primary auditory cortex. In each of the four studies, individuals with schizophrenia and normal controls were pair matched on gender, age at death, and PMI (the latter two matches were as close as possible). Tissue storage time, which was not used in the matching, was also included as a covariate. The same subject matches were used in all four studies and while there were slight differences in the numbers of pairs in each study, there were 15 pairs in common to all four studies that were used in the analyses we present in Section 5.1. In the published analyses for each individual study, a primary model was fit accounting for the effects of subject pairing and tissue storage time on each biomarker to determine the effect of diagnostic group. A secondary model that ignored pairing and instead accounted for the effects of the variables on which subjects were paired, i.e., age at death, gender, and PMI (see Sweet et al. [33][34][35][36] for results) was used to establish robustness of the findings.

Each of the six biomarker measurements was taken in three tissue sections for each subject. However, due to the fact that section numbering is not comparable across studies, the data from each biomarker were averaged across the three sections. This was done by first calculating the mean value for each section and then averaging the three resulting mean values to obtain one average.

Table 5.1: Details of Sweet et al. Auditory Cortical Biomarkers

Biomarker	Brodmann's Area (BA)	Publication	Number of Pairs
Synaptophysin-Immunoreactive (SY-IR) Puncta Density	41	Sweet et al., 2007 [33]	15
Synaptophysin-Immunoreactive (SY-IR) Puncta Density	42		
(Pyramidal Cell) Somal Volume (natural log scale)	41	Sweet et al., 2004 [34]	16
(Pyramidal Cell) Somal Volume (natural log scale)	42	Sweet et al., 2003 [36]	18
(Dendritic) Spine Density	41	Sweet et al., 2008 [35]	15
(Dendritic) Spine Density	42		

5.1.2 Summary of Application Methods for Sweet Studies

5.1.2.1 Linear Discriminant Analysis To account for the effects of subject pairing and tissue storage time on the biomarker data, we applied the pairwise difference approach in Section 3.4.2.2, which produces the same linear discriminant rule as that obtained using the adjustment approaches described in Sections 3.4.2.1 and 3.4.2.3 in the data setting. In addition, we did another analysis ignoring subject pairing and instead adjusted each of the six biomarkers for the effects of age at death, gender, PMI, and storage time, based on the conditional model utilized by Lachenbruch and Tu et al. in Section 3.2.3. The SAS REG procedure was used to adjust the biomarkers using these two methods. For completeness, we also used the approach taken by Knable et al. (2001) and implemented LDA on the original biomarker data, which were not adjusted for either pairing or covariate effects. The classification rules for these three methods were computed using the SAS DISCRIM procedure, assuming equal prior probabilities. In each case, the discriminant coefficients were suitably standardized. We report the observed (resubstituted) misclassification rates in our discussion, along with the misclassification rates obtained using 15-fold cross validation. To implement 15-fold cross validation, we omitted one pair of observations at a time when computing our discriminant rule, which we carried out using a macro we developed in SAS.

5.1.2.2 Classification Trees The construction methodology we developed in Section 4.5.2.2 based on the covariate adjusted feature differences was first applied to the biomarker data to adjust for the effects of subject pairing and tissue storage time. (Note that the adjustment methodology in Section 4.5.2.1 would have yielded a tree with the same structure and the same set of splitting variables, as noted in Section 4.5.2.2). In an additional analysis, we adjusted the six biomarkers for the effects of age at death, gender, PMI, and storage time while ignoring subject pairing, using the semi-parametric tree construction procedure in Section 4.3.3. Finally, we used the approach taken by Knable et al. (2002) and implemented the standard non-parametric BFOS construction procedure as described in Section 4.1.3.2 on the original (unadjusted) biomarker data. The classification trees for these three methods were constructed using Salford Systems CART® software, based on the Gini index and assuming equal prior probabilities. The resulting trees were then pruned using 15-fold cross

validation as described in Section 4.1.3.4.

5.1.3 Results for Sweet Studies

5.1.3.1 Linear Discriminant Analysis For each of the three approaches, we identified those biomarkers with the largest standardized discriminant coefficients (in absolute value) relative to other biomarkers as having the highest discriminatory importance.

After adjusting for the effects of subject pairing and tissue storage time, somal volume and spine density for BA 42, as well as SY-IR puncta density and spine density for BA 41, were identified as the four biomarkers with the highest discriminatory importance. Based on the signs of the coefficients for these four adjusted biomarkers, the following facts hold, assuming all other adjusted biomarkers are held fixed: (1) adjusting for the effect of tissue storage time, SY-IR puncta density for BA 41, as well as somal volume and spine density for BA 42, are larger for normal controls compared with individuals with schizophrenia in a given pair; (2) spine density for BA 41, suitably adjusted for storage time effects, is smaller for normal controls compared with individuals with schizophrenia in a given pair. Using the adjusted linear discriminant rule based on all six biomarkers, we correctly classified 93% of all subjects measured in the data set, while the cross validated correct classification rate for this adjusted rule was 80%.

Alternately, after adjusting for the effects of age at death, gender, PMI, and tissue storage time, SY-IR puncta density and somal volume for BA 41 were identified as the biomarkers that best discriminated between the control and schizophrenia diagnostic groups. Based on the signs of the discriminant coefficients, each of these two adjusted biomarkers are larger for normal controls compared with individuals with schizophrenia, holding all other adjusted biomarker values fixed. Using our adjusted linear discriminant rule, we correctly classified 90% of all subjects, while the cross validated correct classification rate was 77%.

When we applied LDA to the unadjusted biomarker data, SY-IR puncta density and spine density for BA 41 were identified as having the highest discriminatory importance. Based on the signs of the discriminant coefficients, each of these two biomarkers are larger for normal controls compared with individuals with schizophrenia, holding all other biomarker values fixed. Using this linear discriminant rule, which does not adjust for either pairing or

covariate effects, we correctly classified 87% of all subjects, while the cross validated correct classification rate was 77%.

Tables 5.2 and 5.3 provide, respectively, the standardized discriminant coefficients and the classification results for all three approaches.

Table 5.2: Standardized Linear Discriminant Coefficients (Coefficients with relatively large values highlighted in **bold**.)

Approach	Biomarker	BA	Coefficient
Paired LDA with Storage Time	**SY-IR Puncta Density**	**41**	**2.27220**
	SY-IR Puncta Density	42	0.583876
	Somal Volume	41	0.218094
	Somal Volume	**42**	**2.37256**
	Spine Density	**41**	**-2.04184**
	Spine Density	**42**	**2.98220**
Adjusting for Age, PMI Gender, and Storage Time	**SY-IR Puncta Density**	**41**	**1.32272**
	SY-IR Puncta Density	42	-0.471332
	Somal Volume	**41**	**1.20111**
	Somal Volume	42	0.303869
	Spine Density	41	0.499051
	Spine Density	42	0.383178
Unadjusted	**SY-IR Puncta Density**	**41**	**0.831568**
	SY-IR Puncta Density	42	-0.485671
	Somal Volume	41	0.574881
	Somal Volume	42	0.347356
	Spine Density	**41**	**0.919880**
	Spine Density	42	-0.0441287

5.1.3.2 Classification Trees Based on the tree we constructed to account for the effects of subject pairing and tissue storage time, schizophrenia subjects were best discriminated from normal controls by SY-IR puncta density and somal volume for BA 41. Specifically, small values of SY-IR puncta density for BA 41, once adjusted for the effect of tissue storage time, are associated with schizophrenia subjects compared with normal controls in a given pair. Among individuals with large values of this adjusted biomarker in a given pair, small values of somal volume for BA 41, adjusted for the effect of tissue storage time, are associated with schizophrenia subjects, while large values are associated with normal controls. Based on our adjusted tree, 90% of all subjects measured in the data set were correctly classified.

Table 5.3: LDA Classification Results Based on All Six Biomarkers

(C - control, S - schizophrenia, MC Rate - misclassification rate)

Approach	From Diag	Classified as C	Classified as S	Total	Observed MC Rate	15-fold CV MC Rate
Paired LDA with Storage Time	C	14	1	15	**0.07**	**0.2**
	S	1	14	15		
	Total	15	15	30		
Adjusting for Age, PMI, Gender and Storage Time	C	13	2	15	**0.1**	**0.233**
	S	1	14	15		
	Total	14	16	30		
Unadjusted	C	13	2	15	**0.133**	**0.233**
	S	2	11	15		
	Total	16	14	30		

When we ignored subject pairing and instead adjusted the six biomarkers for the effects of age at death, gender, PMI, and tissue storage time, we obtained a tree where individuals with schizophrenia were discriminated from normal controls by SY-IR puncta density, somal volume, and spine density for BA 41. Based on this adjusted tree, small values of adjusted spine density and somal volume for BA 41 are associated with schizophrenia subjects, while large values of adjusted spine density and SY-IR puncta density for BA 41 are associated with normal controls. Among individuals with small adjusted spine densities and large adjusted somal volumes, small adjusted SY-IR puncta density values correspond to schizophrenia subjects while large values correspond to normal controls. In addition, large adjusted spine densities and small adjusted SY-IR puncta densities are associated with schizophrenia subjects. Using our adjusted tree, we correctly classified 90% of all subjects.

When we applied the standard BFOS algorithm to the unadjusted biomarker data, spine density for BA 41 was the only discriminatory biomarker chosen, with low and high values corresponding to schizophrenia subjects and normal controls, respectively. Among the subjects examined, 80% were correctly classified.

Figures 5.1, 5.2, and 5.3 display, respectively, the pruned tree that accounts for pairing and storage time effects, the pruned tree that adjusts for the effects of age at death, gender, PMI, and tissue storage time, and the pruned tree that is based on the unadjusted biomarker data. The classification results corresponding to these three trees are displayed in Table 5.4.

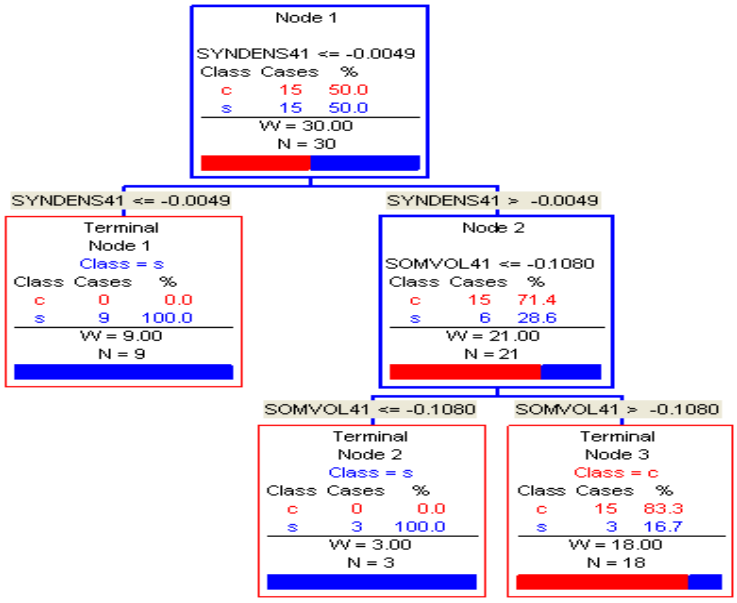

Figure 5.1: Paired classification tree with storage time. SynDens41 and SomVol41 correspond to SY-IR Puncta Density and Somal Volume for BA 41.

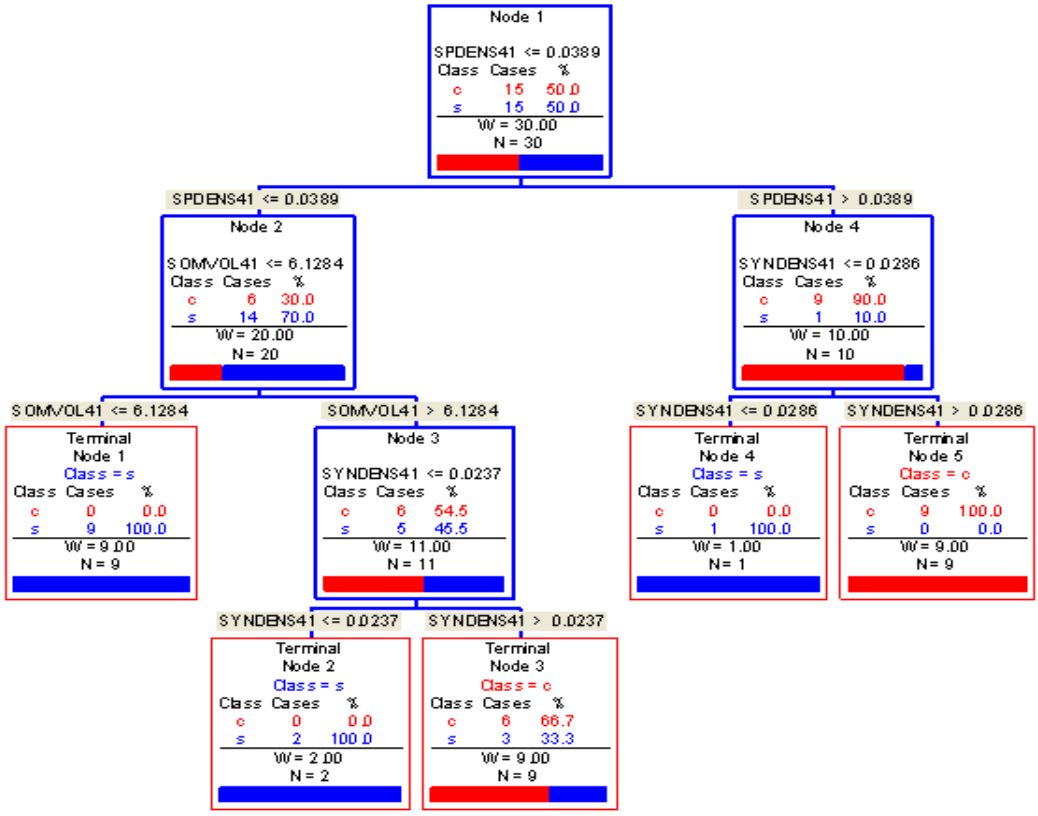

Figure 5.2: Semi-parametric classification tree. SpDens41, SomVol41, and SynDens41 correspond to Spine Density, Somal Volume, and SY-IR Puncta Density for BA 41.

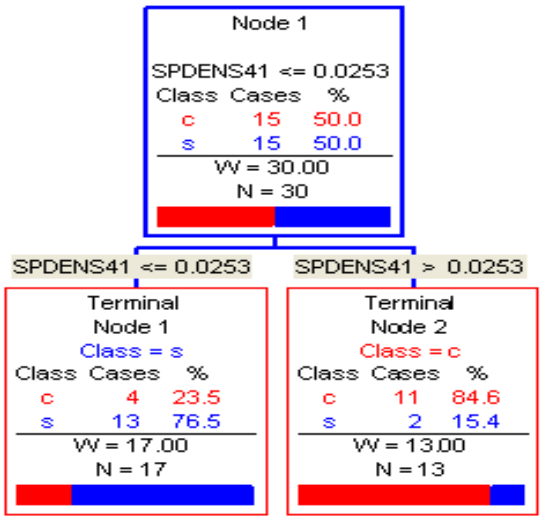

Figure 5.3: Traditional classification tree. SpDens41 corresponds to Spine Density for BA 41.

Table 5.4: Classification Results for Classification Trees Pruned Using 15-fold CV (C - control, S - schizophrenia, MC rate - misclassification rate)

Approach	From Diag	Classified as C	Classified as S	Total	Observed MC Rate
Paired Tree with Storage Time	C S Total	15 3 18	0 12 12	15 15 30	**0.1**
Adjusting for Age, PMI, Gender and Storage Time	C S Total	15 3 18	0 12 12	15 15 30	**0.1**
Unadjusted	C S Total	11 2 13	4 13 17	15 15 30	**0.2**

5.1.4 Discussion of LDA and Classification Tree Results for Sweet Studies

SY-IR puncta density and somal volume for BA 41 consistently appear as important discriminatory biomarkers in our adjusted linear discriminant rules and classification trees. Thus, among all the six biomarkers examined, these seem to be the two biomarkers that best discriminate between the control and schizophrenia diagnostic groups, once we take into account the effects of pairing and brain tissue storage time. However, had we taken the traditional approach to both LDA and classification trees that is seen in the neuroscience literature, as exemplified by Knable et al. (2001, 2002), and not adjusted for either pairing or covariate effects, we would not have been able to identify SY-IR puncta density and somal volume for BA 41 as being the most discriminatory.

From the results in Tables 5.3 and 5.4, we see that the misclassification rates appear to be comparable for the two adjustment methods and the unadjusted approach we use to construct our linear discriminant rules and classification trees. This may be explained by the fact that in earlier models that were fit based on the biomarker data, pairing and tissue storage time effects were not significant for most of the six biomarkers. To elaborate, in each primary model that was fit to account for the effects of subject pairing and tissue storage time on each biomarker, these two effects were only significant for spine density for BA 42 and somal volume for BA 41, respectively. In addition, in each secondary model that was fit to control for the effects of age at death, gender, PMI, and storage time on each biomarker, only the effect of storage time was significant in any of the six models, namely, the model for SY-IR puncta density for BA 41. Based on the model results in this case, adjusting these six biomarkers for the effects of pairing (or the pairing variables) and storage time is not expected to yield substantial gains in classification accuracy.

5.2 KONOPASKE DATA

5.2.1 Description of Dataset

Next, we apply our adjustment methodology to data obtained from a post-mortem brain tissue study conducted by Konopaske et al. [17]. We examine a total of six biomarkers, listed

in Table 5.5, that were measured by Dr. Konopaske and his collaborators in post-mortem monkey brain tissue taken from the left parietal lobe. In this study, 18 male macaque monkeys were matched in triads by terminal body weight where, in each triad, each monkey was treated with a sham drug, haloperidol, or olanzapine, the latter two drugs being antipsychotics. Unlike the Sweet et al. data, no covariates were included in the Konopaske data. To determine the effect of drug treatment group, Konopaske et al. fit an ANOVA model for each biomarker, while controlling for the effect of triad matching (see Konopaske et al. [17] for results). We note that among the examined biomarkers, astrocyte number was the only one that significantly differed between the sham and antipsychotic treatment groups. Thus, our application to the Konopaske data should be viewed as illustrative, rather than providing new insights.

Table 5.5: Details of Konopaske et al. Biomarkers

Biomarker	Publication	Number of Triads
Oligodendrocyte Number	Konopaske et al., 2008 [17]	6
Oligodendrocyte Density		
Ratio of Oligodendrocyte Number to Glial Cell Number (Oligodendrocyte Ratio)		
Astrocyte Number		
Astrocyte Density		
Ratio of Astrocyte Number to Glial Cell Number (Astrocyte Ratio)		

5.2.2 Summary of Application Methods for Konopaske Study

5.2.2.1 Linear Discriminant Analysis To account for the effect of triad matching on the biomarker data, we applied our differencing adjustment method in Section 3.5.2.2 to the six biomarkers, which yields the same results as those obtained using our adjustment procedure in Section 3.5.2.1. For completeness, we also implemented LDA on the origi-

nal biomarker data, which were not adjusted for the effect of matching. In addition, to clearly illustrate the difference between the differencing and stacked adjustment approaches we develop in Sections 3.5.2.2 and 3.5.2.3 in the data setting, we re-applied our differencing approach and also applied our stacked approach to oligodendrocyte number, density, and ratio, where we only included these three biomarkers in our comparison for ease of computation when implementing the stacked approach. Assuming equal priors, we used the SAS DISCRIM procedure to compute the corresponding classification rules. In our discussion, we report the observed misclassification rates, along with those obtained using 6-fold cross validation, where we omitted one triad of observations at a time when computing our linear discriminant rule.

5.2.2.2 Classification Trees We first applied our differencing adjustment approach in Section 4.4.2.2 to the biomarker data. For comparative purposes, we also implemented the standard non-parametric BFOS construction method in Section 4.1.3.2 on the original (unadjusted) biomarker data. To construct these two trees, we used Salford Systems CART® software based on the Gini index and the assumption of equal priors, and pruned them using 6-fold cross validation as discussed in Section 4.1.3.4.

5.2.3 Results for Konopaske Study

5.2.3.1 Linear Discriminant Analysis When we applied each LDA approach, we obtained three linear discriminant functions from which we identified the biomarkers that best discriminate between the haloperidol and olanzapine treatment groups, the haloperidol and sham treatment groups, and the olanzapine and sham treatment groups. We then standardized the coefficients in each discriminant function to identify the biomarkers with the largest standardized discriminant coefficients (in absolute value) relative to the other biomarkers as having the highest discriminatory importance in that discriminant function.

In implementing our differencing adjustment approach and using the signs of the coefficients in each discriminant function, we can state the following when we hold all other biomarker values fixed.

Astrocyte number is largest for sham treated subjects, followed by subjects treated with

haloperidol, and is smallest for olanzapine treated subjects. Also, oligodendrocyte number is larger for olanzapine treated subjects compared with either haloperidol or sham treated subjects, while astrocyte density is smaller for olanzapine treated subjects compared with sham treated subjects. Finally, oligodendrocyte ratios are larger and astrocyte ratios are smaller for sham treated subjects compared with either haloperidol or olanzapine treated subjects. Based on the three estimated linear discriminant functions, we correctly classified 72% of all subjects measured in the data set, while the cross validated correct classification rate was 39%.

Although we leave the details of the results obtained when we apply our differencing and stacked adjustment approaches to oligodendrocyte number, density, and ratio for Appendix D, we point out the fact that these two approaches produced entirely different types of results. Not only do both approaches yield, in their context, different discriminant functions, but also the interpretation of these functions necessarily differs, as explained in Appendix D. These results allow us to see the practical interpretation of the population based difference that we showed in Section 3.5.1.3.

Tables 5.6 and 5.7 provide, respectively, the standardized discriminant coefficients and classification results obtained from implementing the differencing and traditional approach to all six biomarkers.

5.2.3.2 Classification Trees When we applied our differencing adjustment approach, we obtained a tree (see Figure 5.4) where astrocyte number and the ratio of astrocyte number to glial cell number (astrocyte ratio) were identified as the two biomarkers that best discriminated among the three drug treatment groups. To elaborate, once we adjust for the effect of triad matching on the biomarker data, small values of these two discriminatory biomarkers are associated with olanzapine treated subjects. Among subjects with large values of adjusted astrocyte ratio, those with small values of adjusted astrocyte number are associated with the haloperidol treatment group. Regardless of whether adjusted astrocyte ratio is large or small, we have that large values of adjusted astrocyte number correspond to sham treated subjects. Based on this adjusted tree, we correctly classified 83% of all subjects measured in the data set.

Figures 5.4 and 5.5 display, respectively, the matched tree obtained from our differencing

Table 5.6: Standardized Linear Discriminant Coefficients for Haloperidol (H) vs. Olanzapine (O), Haloperidol vs. Sham (S), and Olanzapine vs. Sham

(Coefficients with relatively large values highlighted in **bold**.)

Approach	Biomarker	Coefficient (H vs. O)	Coefficient (H vs. S)	Coefficient (O vs. S)
Matched LDA	Oligodendrocyte Number	**-16.2784**	0.943648	**17.2220**
	Oligodendrocyte Density	3.73358	2.94737	-0.786208
	Oligodendrocyte Ratio	10.9869	**-3.80169**	**-14.7886**
	Astrocyte Number	**22.5686**	**-7.37721**	**-29.9458**
	Astrocyte Density	-10.5711	2.21073	**12.7818**
	Astrocyte Ratio	-8.81832	**4.07302**	**12.8913**
Unadjusted	Oligodendrocyte Number	**-11.7108**	-1.66035	**10.0504**
	Oligodendrocyte Density	2.27314	**4.70486**	2.43173
	Oligodendrocyte Ratio	8.18420	**-3.28950**	**-11.4737**
	Astrocyte Number	**15.1880**	-2.55505	**-17.7431**
	Astrocyte Density	-5.47579	-1.90730	3.56849
	Astrocyte Ratio	-6.90072	**3.71335**	**10.6141**

Table 5.7: LDA Classification Results Based on All Six Biomarkers

(H - haloperidol, O - olanzapine, S - sham, MC rate - misclassification rate)

Approach	From Group	Classified as H	Classified as O	Classified as S	Total	Observed MC rate	6-fold CV MC rate
Matched LDA	H	0	3	3	6	**0.28**	**0.61**
	O	0	6	0	6		
	S	3	2	1	6		
	Total	3	11	4	18		
Unadjusted	H	1	2	3	6	**0.28**	**0.61**
	O	2	4	0	6		
	S	4	0	2	6		
	Total	7	6	5	18		

approach in Section 4.4.2.2 to account for the effect of triad matching on the biomarker data and the tree obtained when we implement the standard BFOS algorithm on the unadjusted biomarker data. The corresponding classification results are provided in Table 5.8.

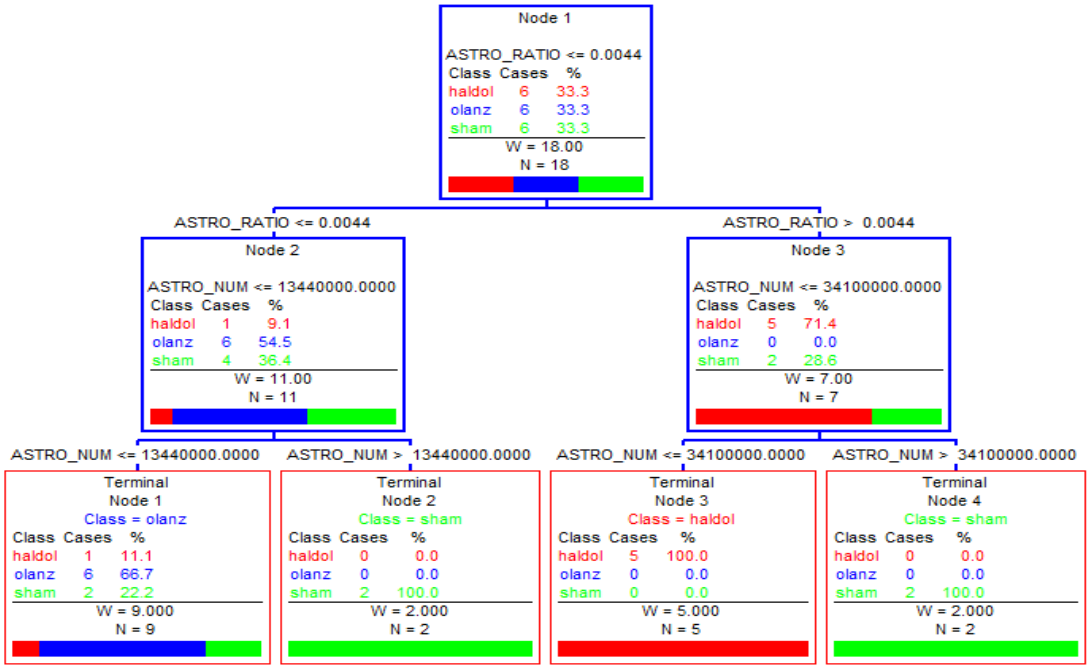

Figure 5.4: Matched classification tree. Astro_Num and Astro_Ratio correspond to astrocyte number and astrocyte ratio.

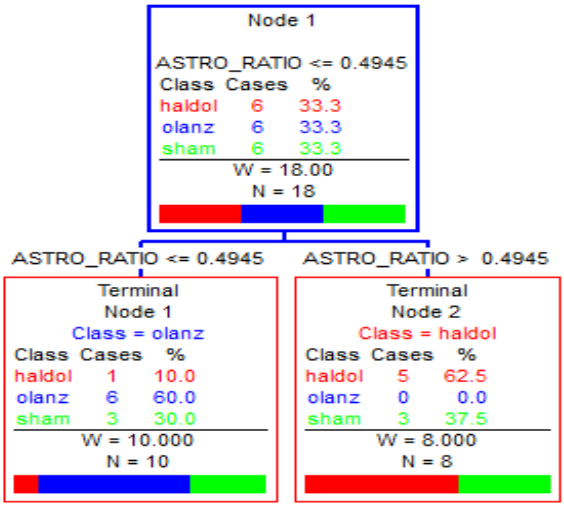

Figure 5.5: Traditional classification tree. Astro_Ratio corresponds to astrocyte ratio.

Table 5.8: Classification Results for Classification Trees Pruned Using 6-fold CV

(H - haloperidol, O - olanzapine, S - sham, MC rate - misclassification rate)

Approach	From Group	Classified as H	Classified as O	Classified as S	Total	Observed MC Rate
Matched Tree	H	5	1	0	6	**0.17**
	O	0	6	0	6	
	S	0	2	4	6	
	Total	5	9	4	18	
Unadjusted	H	5	1	0	6	**0.39**
	O	0	6	0	6	
	S	3	3	0	6	
	Total	8	10	0	18	

5.2.4 Discussion of LDA and Classification Tree Results for Konopaske Study

Regardless of whether we use LDA or classification trees to discriminate among the three treatment groups, astrocyte number and the ratio of astrocyte number to glial cell number are always identified as important discriminatory biomarkers when we adjust for the effect of triad matching. Therefore, once we account for the effect of triad matching on all six biomarkers, we have that these are the two biomarkers that best discriminate among the haloperidol, olanzapine, and sham treatment groups. We note that if we had ignored the effect of triad matching, we would not have identified astrocyte number as a discriminatory biomarker in our classification tree.

Although the classification results obtained from using the differencing and traditional approaches appear to be comparable for LDA, as we see in Table 5.7, we have that our matched classification tree correctly classifies a noticeably higher percentage of the examined subjects, relative to the tree based on the unadjusted biomarker data. Thus, in this instance, our methodology to account for subject matching gives us not only a clearer picture of which of the six biomarkers best discriminate among the three treatment groups, but also more accurate classification results, compared with those obtained when we ignore the effect of matching.

5.3 SUMMARY OF APPLICATION RESULTS

We show that our methodology to adjust for the effects of group matching and covariates in LDA and classification trees can be easily applied to data, e.g., post-mortem tissue data. Based on the results of our application to both the Sweet et al. and Konopaske et al. data, we found that our adjustment methodology allows us to better determine which of the biomarkers in each of these two data sets best discriminates among the diagnostic or treatment groups under consideration and can also yield generally more accurate classification results, compared with traditional LDA or classification trees.

6.0 CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

In this dissertation, we successfully develop a methodology for two commonly used discrimination methods, namely, LDA and classification trees, that adjusts the feature variables of interest for the effects of group matching and covariates. If not properly taken into account, matching and covariates can potentially mask the true discriminatory ability of the feature data. Using our adjustment methodology, we can get a clearer and more accurate picture of which feature variables, among those examined, best discriminate among the g ($g \geq 2$) groups under consideration. In addition, our research methodology for group discrimination can be easily applied to any study where subjects are matched across different groups and/or measured on additional covariates, e.g., post-mortem brain tissue studies.

The concept of adjusting for covariate effects in LDA using the conditional distribution of the feature data was initially explored by Cochran and Bliss [8] and generalized more extensively by Lachenbruch [19] and Tu et al. [37]. However, none of these authors addressed the fact that individuals may be matched across the g groups under consideration and that such matching may also greatly impact the feature variables under study. Therefore, an extension of these authors' covariate adjustment methodologies to also account for the effect of group matching is clearly required. In Chapter 3, we successfully formulate and develop an extension of these authors' covariate adjustment methodologies to also account for the effect of group matching in both a theoretical framework and in the context of data.

On the other hand, there appears to be little in the literature that deals with accounting for either the effects of group matching or covariates on the feature variables of interest when constructing classification trees. In the spirit of Lachenbruch and Tu et al., we carefully

develop in Chapter 4 a tree construction methodology that adjusts for these two effects by incorporating the conditional distribution of the examined feature variables for a given match and/or a given set of covariate values. We begin by detailing how the BFOS tree construction method can be implemented in the case where the feature data belong to a known continuous distribution in each group. This approach subsequently leads to our development of a parametric alternative to the standard non-parametric BFOS algorithm that can be implemented using training data. More importantly, this population based approach provides us with a basis to use the conditional distribution of the feature data when constructing a particular tree. We then formulate a specific semi-parametric model for the conditional distribution of the feature data which allows us to construct a tree that suitably adjusts for covariate effects and is based on a unique set of feature variables that does not change depending on the value at which a set of covariates is fixed. In our development, we show that our semi-parametric tree construction procedure can be easily applied to training data by using standard classification and regression tree software packages. Thus, there is no need to develop new software to implement our approach. An extension of our semi-parametric construction methodology is developed for the case where individuals are matched across two or more groups, and then to the case in which matched individuals are measured on additional covariates.

In Chapter 5, we successfully apply our adjustment methodology for LDA and classification trees to two post-mortem brain tissue data sets in which subjects are matched, namely, that obtained from the studies conducted by Sweet et al. [33][34][35][36] and that from the study conducted by Konopaske et al. [17]. In applying our methodology to the Sweet data, we are able to identify among the six biomarkers of interest those that best distinguish a control subject from a schizophrenia subject in any given pair, while, at the same time, accounting for the effect of brain tissue storage time on these biomarkers. Also, when we apply our adjustment procedure to the Konopaske data, we can determine the biomarkers that best discriminate among the sham, haloperidol, and olanzapine treatment groups, while taking into account the effect of triad matching.

6.2 FUTURE WORK

6.2.1 Discriminant Analysis

As we noted in Section 3.2.3, the methodology behind general covariance adjusted LDA can be easily extended to quadratic discriminant analysis (QDA). To elaborate, Tu et al. [37] also develop general covariance adjusted QDA by considering the case in which given $\mathbf{X} = \mathbf{x}$, $\mathbf{Y} \sim N_P(h_i(\mathbf{x}), \boldsymbol{\Sigma}_i)$ in the i^{th} group $(i = 1, \ldots, g)$, where $\boldsymbol{\Sigma}_i$ is the conditional variance-covariance matrix in the i^{th} group, $h_i(\mathbf{x}) = \boldsymbol{\mu}_i + \boldsymbol{\rho}_i(\mathbf{x}; \boldsymbol{\Theta}_i)$, $\boldsymbol{\mu}_i$ corresponds to the effect of the i^{th} group on $\mathbf{Y}$, and $\boldsymbol{\rho}_i(\mathbf{x}; \boldsymbol{\Theta}_i) = (\rho_{1,i}(\mathbf{x}; \boldsymbol{\theta}_{1,i}), \ldots, \rho_{P,i}(\mathbf{x}; \boldsymbol{\theta}_{P,i}))'$ is a known smooth function of $\mathbf{x}$ and the parameter vectors $\boldsymbol{\theta}_{1,i}, \ldots, \boldsymbol{\theta}_{P,i}$ in the i^{th} group. Since Tu et al. do not address in their development the fact that individuals may also be matched, we would like to extend general covariance adjusted QDA to also account for the effect of group matching on the feature variables of interest.

If the number of elements in $\mathbf{Y} = (Y_1, \ldots, Y_P)'$ is smaller than the number of observations in the training data, N, then no major issues arise when we implement either LDA or QDA based on the training data. However, if we have high dimensional feature data ($P >> N$), e.g., microarray data, then we encounter a number of issues, including the fact that the parameter estimates used in traditional LDA and QDA may be highly unstable [11]. To address these issues, Friedman proposed regularized discriminant analysis (RDA), a "hybrid" between traditional LDA and QDA, which shrinks the group specific variance-covariance matrices in QDA to one that is common to all groups, as in LDA [11][14]. Specifically, the estimated regularized variance-covariance matrix in the i^{th} group $(i = 1, \ldots, g)$ has the form $\hat{\boldsymbol{\Sigma}}_{YY,i}(\eta) = \eta\hat{\boldsymbol{\Sigma}}_{YY,i} + (1-\eta)\hat{\boldsymbol{\Sigma}}_{YY}$, where $\eta \in [0, 1]$, and $\hat{\boldsymbol{\Sigma}}_{YY,i}$ and $\hat{\boldsymbol{\Sigma}}_{YY}$ are the group specific and pooled estimates of the variance-covariance matrix of $\mathbf{Y}$. We would like to explore how RDA can be modified to account for the effects of group matching and covariates on the feature data.

Tu et al. [37] also introduce general covariance adjusted logistic discriminant analysis for two groups. Specifically, Tu et al. define the covariance adjusted logistic discriminant function as

$$\log\left[\frac{f_1(\mathbf{y}|\mathbf{x})}{f_2(\mathbf{y}|\mathbf{x})}\right] = \eta + \boldsymbol{\varsigma}\tilde{\mathbf{y}},$$

for a given $\mathbf{x}$, where the conditional densities $f_i(\mathbf{y}|\mathbf{x})$ $(i = 1, 2)$ are defined as in Section 3.2.1.1, $\eta \in \mathbb{R}$, $\boldsymbol{\varsigma} = (\varsigma_1, \ldots, \varsigma_P)$, $\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$ denotes the covariate adjusted feature vector, and $\boldsymbol{\rho}(\mathbf{x}; \boldsymbol{\Theta})$ is defined as in Section 3.2.3. In this modification to traditional logistic discriminant analysis, the feature data are only adjusted for covariate effects. Due to the fact that individuals may also be paired across the two groups, we are interested in extending general covariance adjusted logistic discriminant analysis to also handle subject pairing.

6.2.2 Classification Trees

Our assumption of equal misclassification costs is prevalent in all of the adjustment methodologies we develop in Sections 4.2 to 4.5 for classification trees. In certain contexts, this assumption may not be appropriate and, thus, we are interested in exploring ways to modify these construction procedures to handle variable misclassification costs. However, additional issues may arise in using this particular cost structure. For example, the Gini index is no longer necessarily a strictly concave function of $P(\mathbf{Y} \in \text{group } i|\mathbf{Y} \in t)$ under this cost structure and, thus, it is possible that the impurity of node t may increase when it is split in this instance [7].

In Sections 4.1.2.3 and 4.2.2.2, we developed two tree construction approaches for normal populations based on the same assumptions made in traditional and covariance adjusted LDA, respectively. We also noted the parallel between the classification trees using these two approaches and the classification regions obtained from traditional and covariance adjusted LDA in the case of univariate $\mathbf{Y}$. However, we would like to further extend this comparison of classification trees for normal data and linear discriminant classification regions for multivariate $\mathbf{Y}$. For instance, we would like to theoretically justify whether traditional or covariance adjusted LDA are better methods to partition the feature space $\mathcal{Y}$ than the construction methods we formulate in Sections 4.1.2.3 and 4.2.2.2, respectively, or vice versa. We are also interested in exploring how the BFOS tree construction algorithm can be implemented for normal populations based on the assumptions made in traditional and covariance adjusted QDA, as well as how the trees obtained from these two approaches compare with the classification regions obtained from traditional and covariance adjusted QDA.

If we're dealing with high dimensional feature data, the same issues arise when we imple-

ment the BFOS algorithm for normal data as those that arose for traditional LDA and QDA. In this case, we would like to explore whether this normal-based tree construction method can be modified by regularizing the variance-covariance matrices of $\mathbf{Y}$ using the approach Friedman took in his development of RDA. We are also interested in exploring whether we can implement this modification if we consider the conditional distribution of the normal feature data for a given covariate value. Assuming we can modify our traditional and conditional data-based tree construction procedures in Sections 4.1.3.1 and 4.2.3.1, respectively, in such a manner, we would like to explore how the corresponding results compare with those obtained from traditional RDA, as well as RDA where the feature data have been suitably adjusted for covariate effects.

In Section 4.1.1, we noted that classification trees can also split the feature space using linear combinations of the feature variables in $\mathbf{Y}$ [7]. Such splits partition the feature space using a particular set of hyperplanes, as is the case for traditional LDA. Although this construction approach yields trees that are considerably harder to interpret, relative to trees obtained using splits of the form $Y \leq c$, we are interested in examining how traditional classification trees using linear combination splits compare with traditional LDA when we deal with normal populations. We would also like to explore how our conditional tree construction approach in Section 4.2.2.2 for normal populations can be modified to handle linear combination splits, and how the trees obtained from this approach compare with the classification regions in covariance adjusted LDA.

To further illustrate the improvement of our semi-parametric and matched classification tree methodologies, we may consider implementing them using simulated data, in order to examine how these resulting trees compare with trees constructed using the traditional BFOS algorithm. We are also interested in investigating other estimation methods beyond that of LS estimation that can be used to estimate the parameters needed to construct our semi-parametric trees, as well as our matched classification trees.

6.2.3 Tree Ensemble Construction Methods

Although there are several benefits to using classification trees, not least of which is their simple and interpretable structure, they all share one important weakness. They are consid-

ered unstable because relatively small changes to the training data may lead to large changes in the resulting tree [3]. Therefore, various authors have looked into possible solutions to this issue, the most well known of which is Breiman's, who proved that the accuracy of all types of unstable classifiers could be increased by generating multiple classifiers obtained by permuting the training data set or construction method and aggregating them to yield one single classifier or ensemble [3][4]. He then proceeded to develop the bootstrap aggregating or bagging algorithm [3], in which B classification trees are constructed using B bootstrap replicates of the original training data set. Unlike the BFOS algorithm, trees constructed using the bagging algorithm are not pruned. With the original training data used as a test set, each individual is finally assigned to the group having the majority among the B trees. Several extensions to bagging have subsequently been developed, the most notable of which is the random forests algorithm [6].

As with bagging, random forests consist of B classification trees constructed from B bootstrap replicates of the original training data set. However, for random forests, only a randomly chosen subset of the P feature variables in $\mathbf{Y}$ is considered when constructing each of these B trees. Each tree is then used to classify all individuals in the training data so that each individual is classified into the majority group among all trees in a particular forest. To clarify, suppose we construct a random forest of 100 trees based on the Sweet et al. biomarker data. If a subject is assigned to the control diagnostic group in 60 of the trees and assigned to the schizophrenia diagnostic group in the remaining 40 trees, then the random forest based on the Sweet et al. data would classify this subject into the control group.

However, we note that each classification tree in a random forest or similar tree ensemble is constructed using the same procedure utilized in the traditional BFOS recursive partitioning algorithm. Therefore, we can easily apply to random forests the adjustment methodologies we develop for semi-parametric trees and matched classification trees in Sections 4.3 to 4.5 to account for the effects of group matching and/or covariates, an application we plan to refine further post-dissertation.

6.2.4 Clustering

In the context of post-mortem tissue studies, schizophrenia has been considered a disease consisting of various subtypes. Due to the fact that this heterogeneity may be explained by examining select brain regions, another goal of such studies is to analyze a particular set of biomarkers in order to identify possible subpopulations of subjects with schizophrenia. However, when determining these clusters of schizophrenia subjects, it is important to account for the fact that subjects in these studies are paired and also measured on additional covariates, as was addressed by Wu in his development of several methods to cluster subjects with schizophrenia in post-mortem tissue studies [40]. Although it is beyond the scope of this dissertation, we would like to examine how the adjustment methodology we develop for LDA and classification trees can be applied to various clustering techniques, so that we may adjust for matching and covariate effects when we use clustering to reveal subpopulations that may exist among a group of individuals.

6.3 SUMMARY

In studies where individuals from different groups are measured on a particular set of feature variables, it may be of interest to determine which of these variables best discriminate among these groups. When these individuals are also matched across these groups and measured on additional covariates, it is important to account for both matching and covariate effects when determining the discriminatory ability of the feature variables of interest in order to avoid obtaining misleading results. However, there appears to be nothing in the literature that incorporates both subject matching and covariate effects in the implementation of any discrimination procedure, including LDA and classification trees. Due to their fairly common usage, our research concentrates on modifying these two discriminatory methods to adjust for the effects of group matching and covariates on the feature data. For any study that involves matching subjects across different groups and/or measuring these subjects on other covariates, the research methodology we develop in this dissertation is highly beneficial and has potentially powerful applications.

APPENDIX A

A.1 CLASSIFYING TWO PAIR MEMBERS IN LDA USING KNOWN PAIR EFFECT

For a given pair, we assume $\mathbf{Y}^+ \sim N_{2P}(\boldsymbol{\mu}_i^+, \boldsymbol{\Sigma}^+)$ in the i^{th} group ordering $(i = 1, 2)$ of $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$, where

$$\boldsymbol{\mu}_1^+ = \begin{bmatrix} \boldsymbol{\mu}_1 + \boldsymbol{\gamma} \\ \boldsymbol{\mu}_2 + \boldsymbol{\gamma} \end{bmatrix}, \qquad \boldsymbol{\mu}_2^+ = \begin{bmatrix} \boldsymbol{\mu}_2 + \boldsymbol{\gamma} \\ \boldsymbol{\mu}_1 + \boldsymbol{\gamma} \end{bmatrix}, \qquad \boldsymbol{\Sigma}^+ = \begin{bmatrix} \boldsymbol{\Sigma} + \boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Sigma} + \boldsymbol{\Psi} \end{bmatrix},$$

and $\boldsymbol{\Psi}$ represents the covariance between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$. To ensure that $\boldsymbol{\Sigma}^+$ is positive definite in this instance, we assume that both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} + 2\boldsymbol{\Psi}$ are positive definite.

Due to the invariance of LDA to nonsingular transformations [1][23], $\mathbf{Y}^+$ and $\mathbf{A}\mathbf{Y}^+$ yield the same linear discriminant classification rule, where $\mathbf{A}$ is a nonsingular $2P \times 2P$ matrix. For our purposes, we let $\mathbf{A} = \left[\begin{smallmatrix} \mathbf{I}_P & -\mathbf{I}_P \\ \mathbf{I}_P & \mathbf{I}_P \end{smallmatrix}\right]$, where $\mathbf{I}_P$ is the $P \times P$ identity matrix, such that $\mathbf{A}\mathbf{Y}^+ = \left[\begin{smallmatrix} \mathbf{Y}_{ind} - \mathbf{Y}_{sib} \\ \mathbf{Y}_{ind} + \mathbf{Y}_{sib} \end{smallmatrix}\right] \equiv \left[\begin{smallmatrix} \mathbf{U} \\ \mathbf{V} \end{smallmatrix}\right]$. We then have that $\mathbf{A}\mathbf{Y}^+ \sim N_{2P}(\mathbf{A}\boldsymbol{\mu}_i^+, \mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}')$ in the i^{th} group ordering $(i = 1, 2)$, where

$$\mathbf{A}\boldsymbol{\mu}_1^+ = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 + 2\boldsymbol{\gamma} \end{bmatrix}, \qquad \mathbf{A}\boldsymbol{\mu}_2^+ = \begin{bmatrix} -\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 + 2\boldsymbol{\gamma} \end{bmatrix}, \qquad \mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}' = \begin{bmatrix} 2\boldsymbol{\Sigma}_* & \mathbf{0} \\ \mathbf{0} & 2\boldsymbol{\Sigma}_1 \end{bmatrix},$$

$\boldsymbol{\eta}_1 = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\eta}_2 = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma} + 2\boldsymbol{\Psi}$.

From this parametrization, we have that $\boldsymbol{\Sigma}_*$ and $\boldsymbol{\Sigma}_1$ are both positive definite based on our previous assumption regarding $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} + 2\boldsymbol{\Psi}$. Based on our assumptions regarding $\boldsymbol{\Sigma}_*$ and $\boldsymbol{\Sigma}_1$, it follows that $\mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}'$ is also positive definite. Since the matrix $\mathbf{W}$ is positive definite if and only if $\mathbf{B}\mathbf{W}\mathbf{B}'$ is positive definite, where $\mathbf{B}$ is a nonsingular square matrix, $\mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}'$ being positive definite implies that $(\mathbf{A}^{-1})(\mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}')(\mathbf{A}^{-1})' = \boldsymbol{\Sigma}^+$ is also positive definite. We note here that our assumption that $\boldsymbol{\Psi}' = \boldsymbol{\Psi}$ is a sufficient condition for $\mathbf{U}$ and $\mathbf{V}$ to be independent.

Since $\mathbf{U}$ and $\mathbf{V}$ are independent, and the distribution of $\mathbf{V}$ provides no information for discriminatory purposes, $\mathbf{V}$ can be ignored when constructing our classification rule. In other words, we only need to consider the densities of $\mathbf{U}$ in each group ordering. As we stated in Section 3.3.1, it is equally likely that each member of a given pair belongs to either of the two groups and the labeling of a member as an individual or a sibling is assumed to be completely random. Thus, we assume that the prior probability of each group ordering is 0.5. We can then apply to $\mathbf{U}$ the rule in (3.4) used for traditional LDA in order to obtain the rule

$$R_1^+ : \ (\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \boldsymbol{\Sigma}_*^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \quad R_2^+ : \ (\mathbf{y}_{ind} - \mathbf{y}_{sib})' \, \boldsymbol{\Sigma}_*^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0,$$

which is identical to the rule in (3.16). In particular, we classify an individual in a pair and their sibling into the first and second groups, respectively, if $\mathbf{y}_{ind} - \mathbf{y}_{sib}$ falls into region R_1^+, and vice versa if $\mathbf{y}_{ind} - \mathbf{y}_{sib}$ falls into region R_2^+.

A.2 CLASSIFYING TWO PAIR MEMBERS IN LDA USING KNOWN PAIR AND COVARIATE EFFECTS

Given $\mathbf{X}^+ = \mathbf{x}^+$, we assume $\mathbf{Y}^+ \sim N_{2P}(\boldsymbol{\mu}_{i(x)}^+, \boldsymbol{\Sigma}_{(x)}^+)$ in the i^{th} group ordering $(i = 1, 2)$ of $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$ for a given pair, where

$$\boldsymbol{\mu}_{1(x)}^+ = \begin{bmatrix} \boldsymbol{\mu}_1 + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_2 + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{sib} \end{bmatrix}, \quad \boldsymbol{\mu}_{2(x)}^+ = \begin{bmatrix} \boldsymbol{\mu}_2 + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_1 + \boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{sib} \end{bmatrix}, \quad \boldsymbol{\Sigma}_{(x)}^+ = \begin{bmatrix} \boldsymbol{\Sigma}_{(x)} + \boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{(x)} + \boldsymbol{\Psi} \end{bmatrix},$$

and $\boldsymbol{\Psi}$ represents the covariance between $\mathbf{Y}_{ind}$ and $\mathbf{Y}_{sib}$. To ensure that $\boldsymbol{\Sigma}_{(x)}^+$ is positive definite, we assume that both $\boldsymbol{\Sigma}_{(x)}$ and $\boldsymbol{\Sigma}_{(x)} + 2\boldsymbol{\Psi}$ are positive definite.

Based on the invariance property of LDA, we can equivalently consider the conditional distribution of $\mathbf{AY}^+$, where $\mathbf{A}$ and $\mathbf{AY}^+ \equiv \left[\begin{smallmatrix} \mathbf{U} \\ \mathbf{V} \end{smallmatrix}\right]$ are defined as in Appendix A.1. Specifically, given $\mathbf{X}^+ = \mathbf{x}^+$, $\mathbf{AY}^+ \sim N_{2P}(\mathbf{A}\boldsymbol{\mu}_{i(x)}^+, \mathbf{A}\boldsymbol{\Sigma}_{(x)}^+\mathbf{A}')$ in the i^{th} group ordering $(i = 1, 2)$, where

$$\mathbf{A}\boldsymbol{\mu}_{1(x)}^+ = \begin{bmatrix} \boldsymbol{\eta}_1 + \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib}) \\ \boldsymbol{\eta}_2 + 2\boldsymbol{\gamma} + \boldsymbol{\beta}(\mathbf{x}_{ind} + \mathbf{x}_{sib}) \end{bmatrix}, \ \mathbf{A}\boldsymbol{\mu}_{2(x)}^+ = \begin{bmatrix} -\boldsymbol{\eta}_1 + \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib}) \\ \boldsymbol{\eta}_2 + 2\boldsymbol{\gamma} + \boldsymbol{\beta}(\mathbf{x}_{ind} + \mathbf{x}_{sib}) \end{bmatrix}, \ \mathbf{A}\boldsymbol{\Sigma}_{(x)}^+\mathbf{A}' = \begin{bmatrix} 2\boldsymbol{\Sigma}_{*(x)} & \mathbf{0} \\ \mathbf{0} & 2\boldsymbol{\Sigma}_{1(x)} \end{bmatrix},$$

$\boldsymbol{\Sigma}_{*(x)} = \boldsymbol{\Sigma}_{(x)}$, $\boldsymbol{\Sigma}_{1(x)} = \boldsymbol{\Sigma}_{(x)} + 2\boldsymbol{\Psi}$, and $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are defined as in Appendix A.1.

Based on our previous assumption regarding $\boldsymbol{\Sigma}_{(x)}$ and $\boldsymbol{\Sigma}_{(x)} + 2\boldsymbol{\Psi}$, the matrices $\boldsymbol{\Sigma}_{*(x)}$ and $\boldsymbol{\Sigma}_{1(x)}$ are both positive definite and, thus, $\mathbf{A}\boldsymbol{\Sigma}^{+}_{(x)}\mathbf{A}'$ is also positive definite. Using the same argument as in Appendix A.1, we have that if $\mathbf{A}\boldsymbol{\Sigma}^{+}_{(x)}\mathbf{A}'$ is positive definite, then $\boldsymbol{\Sigma}^{+}_{(x)}$ is also positive definite. Our assumption that the covariance matrix $\boldsymbol{\Psi}$ is symmetric is a sufficient condition for $\mathbf{U} \equiv \mathbf{Y}_{ind} - \mathbf{Y}_{sib}$ and $\mathbf{V} \equiv \mathbf{Y}_{ind} + \mathbf{Y}_{sib}$ to be independent.

For a given value of $\mathbf{x}^{+}$, $\mathbf{U}$ and $\mathbf{V}$ are independent, and the distribution of $\mathbf{V}$ provides no information that is useful for discrimination purposes. Therefore, we can ignore $\mathbf{V}$ and only consider the conditional distribution of $\mathbf{U}$ in each group ordering. Retaining our assumption from Appendix A.1 of equal priors for each group ordering, we can apply general covariance adjusted LDA based on the conditional distributions of $\mathbf{U}$ in each group ordering to obtain the rule

$$\begin{aligned} R^{+}_{1(x)} &: \ [(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})]' \, \boldsymbol{\Sigma}^{-1}_{*(x)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0, \\ R^{+}_{2(x)} &: \ [(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})]' \, \boldsymbol{\Sigma}^{-1}_{*(x)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < 0, \end{aligned} \tag{A.1}$$

which is identical to the rule provided in (3.22). Based on (A.1), we classify an individual in a pair and his or her sibling into the first and second groups, respectively, if $(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})$ falls into region $R^{+}_{1(x)}$, and vice versa if $(\mathbf{y}_{ind} - \mathbf{y}_{sib}) - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib})$ falls into region $R^{+}_{2(x)}$.

We note that even if $\boldsymbol{\beta}$ differs between the two groups, we can still implement our discriminant approach based on $\mathbf{Y}^{+}$. In this case, however, it can be shown that $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^{-1}_{*(x)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ in (A.1) are now replaced with $\frac{1}{2}(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)$ and $\boldsymbol{\Sigma}^{-1}_{*(x)}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)(\mathbf{x}_{ind} - \mathbf{x}_{sib})\right]$, respectively. Therefore, if the (assumed linear) relationship between the feature and covariate data depends on group, then the subset of covariate adjusted feature variables we identify as best discriminating between an individual in group 1 and an individual in group 2 in a given pair also depends on the covariate difference $\mathbf{x}_{ind} - \mathbf{x}_{sib}$.

APPENDIX B

B.1 CLASSIFYING ALL MEMBERS OF A MATCH IN LDA USING KNOWN MATCH EFFECT

For notational convenience, we consider the case of matching across three groups, where $\mathbf{Y}^+ \sim N_{3P}(\boldsymbol{\mu}_i^+, \boldsymbol{\Sigma}^+)$ in the i^{th} group ordering $(i = 1, \ldots, 6)$ of $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, and $\mathbf{Y}_{sib,2}$ for a given match, where

$$\boldsymbol{\mu}_1^+ = \begin{bmatrix} \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_2^+ = \begin{bmatrix} \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_3^+ = \begin{bmatrix} \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_4^+ = \begin{bmatrix} \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \end{bmatrix},$$
$$\boldsymbol{\mu}_5^+ = \begin{bmatrix} \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\mu}_6^+ = \begin{bmatrix} \boldsymbol{\mu}_3+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma} \end{bmatrix}, \quad \boldsymbol{\Sigma}^+ = \begin{bmatrix} \boldsymbol{\Sigma}+\boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}+\boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Sigma}+\boldsymbol{\Psi} \end{bmatrix},$$

and $\boldsymbol{\Psi}$ represents the covariance between any two of the three random feature vectors in that match such that $\boldsymbol{\Psi}' = \boldsymbol{\Psi}$. We assume $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} + 3\boldsymbol{\Psi}$ are positive definite to ensure that $\boldsymbol{\Sigma}^+$ is positive definite.

Since LDA is invariant to nonsingular transformations, constructing a classification rule using either $\mathbf{Y}^+$ or $\mathbf{A}\mathbf{Y}^+$ yields the same result, where $\mathbf{A}$ is a nonsingular $3P \times 3P$ matrix in this case. For our purposes, we let $\mathbf{A} = \begin{bmatrix} \mathbf{I}_P & -\frac{1}{2}\mathbf{I}_P & -\frac{1}{2}\mathbf{I}_P \\ \mathbf{0} & \mathbf{I}_P & \mathbf{I}_P \\ \mathbf{I}_P & \mathbf{I}_P & \mathbf{I}_P \end{bmatrix}$ so that $\mathbf{A}\mathbf{Y}^+ = \begin{bmatrix} \mathbf{Y}_{ind} - \frac{1}{2}(\mathbf{Y}_{sib,1}+\mathbf{Y}_{sib,2}) \\ \mathbf{Y}_{sib,1} - \mathbf{Y}_{sib,2} \\ \mathbf{Y}_{ind}+\mathbf{Y}_{sib,1}+\mathbf{Y}_{sib,2} \end{bmatrix}$ $\equiv \begin{bmatrix} \mathbf{Z} \\ \mathbf{S} \\ \mathbf{W} \end{bmatrix}$. Thus, $\mathbf{A}\mathbf{Y}^+ \sim N_{3P}(\mathbf{A}\boldsymbol{\mu}_i^+, \mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}')$ in the i^{th} group ordering $(i = 1, \ldots, 6)$, where

$$\mathbf{A}\boldsymbol{\mu}_1^+ = \begin{bmatrix} \boldsymbol{\upsilon}_1 \\ \boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_3+3\boldsymbol{\gamma} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}_2^+ = \begin{bmatrix} \boldsymbol{\upsilon}_1 \\ -\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_3+3\boldsymbol{\gamma} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}_3^+ = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1+\frac{3}{4}\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_1+\frac{1}{2}\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_3+3\boldsymbol{\gamma} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}_4^+ = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1+\frac{3}{4}\boldsymbol{\upsilon}_2 \\ -\boldsymbol{\upsilon}_1-\frac{1}{2}\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_3+3\boldsymbol{\gamma} \end{bmatrix},$$
$$\mathbf{A}\boldsymbol{\mu}_5^+ = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1-\frac{3}{4}\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_1-\frac{1}{2}\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_3+3\boldsymbol{\gamma} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}_6^+ = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1-\frac{3}{4}\boldsymbol{\upsilon}_2 \\ -\boldsymbol{\upsilon}_1+\frac{1}{2}\boldsymbol{\upsilon}_2 \\ \boldsymbol{\upsilon}_3+3\boldsymbol{\gamma} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}' = \begin{bmatrix} \frac{3}{2}\boldsymbol{\Sigma}_* & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\boldsymbol{\Sigma}_* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 3\boldsymbol{\Sigma}_1 \end{bmatrix},$$

$\boldsymbol{\upsilon}_1 = \boldsymbol{\mu}_1 - \frac{1}{2}(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_3)$, $\boldsymbol{\upsilon}_2 = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$, $\boldsymbol{\upsilon}_3 = \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2 + \boldsymbol{\mu}_3$, $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma} + 3\boldsymbol{\Psi}$. Based on this parametrization, the parameters $\boldsymbol{\upsilon}_1$, $\boldsymbol{\upsilon}_2$, and $\boldsymbol{\upsilon}_3$ are all in $\mathbb{R}^P$. Due to the fact that $\boldsymbol{\Sigma}_*$ and $\boldsymbol{\Sigma}_1$ are positive definite based on our previous assumption regarding $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} + 3\boldsymbol{\Psi}$, the covariance matrix $\mathbf{A}\boldsymbol{\Sigma}^+\mathbf{A}'$ is also positive definite and, thus, so is $\boldsymbol{\Sigma}^+$. It is easy to show

that our assumption that $\boldsymbol{\Psi} = \boldsymbol{\Psi}'$ is a sufficient condition for the random vectors $\mathbf{Z}$, $\mathbf{S}$, and $\mathbf{W}$ to be mutually independent.

Since $\mathbf{W}$ is independent from both $\mathbf{Z}$ and $\mathbf{S}$ and the distribution of $\mathbf{W}$ provides no discriminatory information, $\mathbf{W}$ can be ignored so that we only consider the distribution of $\left[\begin{smallmatrix}\mathbf{Z}\\ \mathbf{S}\end{smallmatrix}\right]$ in each group ordering when constructing our classification rule. Since each triad member is equally likely to belong to one of the three groups, as was discussed in Section 3.5.1, and the labeling of a member as an individual or as any of the other two siblings is completely random, we assume that each group ordering is equally likely. In other words, we assume the prior probability of each group ordering is 1/6. We can then apply traditional LDA for multiple groups to the stacked differenced vector $\left[\begin{smallmatrix}\mathbf{Z}\\ \mathbf{S}\end{smallmatrix}\right]$, which yields the following classification regions:

$$R_i^+ : \left[\mathbf{d}^* - \frac{1}{2}\left(\boldsymbol{\mu}_i^{+(*)} + \boldsymbol{\mu}_j^{+(*)}\right)\right]' \begin{bmatrix} \frac{2}{3}\boldsymbol{\Sigma}_*^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\boldsymbol{\Sigma}_*^{-1} \end{bmatrix} \left(\boldsymbol{\mu}_i^{+(*)} - \boldsymbol{\mu}_j^{+(*)}\right) > 0 \quad j = 1, \ldots, 6;\ j \neq i, \tag{B.1}$$

where $\mathbf{d}^*$ is the observed counterpart of $\left[\begin{smallmatrix}\mathbf{Z}\\ \mathbf{S}\end{smallmatrix}\right]$ and $\boldsymbol{\mu}_i^{+(*)}$ denotes the first $2P$ components of $\mathbf{A}\boldsymbol{\mu}_i^+$ $(i = 1, \ldots, 6)$. From (B.1), we have that a new triad with an observed feature value $\mathbf{d}^*$ would be classified into the i^{th} group ordering if $\mathbf{d}^*$ falls into region R_i^+ $(i = 1, \ldots, 6)$.

Using the formulas for $\boldsymbol{\mu}_i^{+(*)}$ and $\boldsymbol{\mu}_j^{+(*)}$, we can re-express the rule in (B.1) as

$$R_i^+ : d_{ij} > 0 \quad j = 1, \ldots, 6;\ j \neq i, \tag{B.2}$$

where

$$\begin{aligned}
d_{12} &= (\mathbf{y}_{sib,1} - \mathbf{y}_{sib,2})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3), \quad d_{13} = (\mathbf{y}_{ind} - \mathbf{y}_{sib,1})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),\\
d_{16} &= (\mathbf{y}_{ind} - \mathbf{y}_{sib,2})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3), \quad d_{24} = (\mathbf{y}_{ind} - \mathbf{y}_{sib,2})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),\\
d_{25} &= (\mathbf{y}_{ind} - \mathbf{y}_{sib,1})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3), \quad d_{34} = (\mathbf{y}_{sib,1} - \mathbf{y}_{sib,2})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3),\\
d_{35} &= (\mathbf{y}_{ind} - \mathbf{y}_{sib,2})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3), \quad d_{46} = (\mathbf{y}_{ind} - \mathbf{y}_{sib,1})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3),\\
d_{56} &= (\mathbf{y}_{sib,1} - \mathbf{y}_{sib,2})' \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),
\end{aligned}$$

$d_{14} = d_{12} + d_{24}$, $d_{15} = d_{13} + d_{35}$, $d_{23} = d_{13} - d_{12}$, $d_{26} = d_{24} + d_{46}$, $d_{36} = d_{16} - d_{13}$, and $d_{45} = d_{25} - d_{24}$. We can compute the other 15 d_{ij} functions by using the fact that $d_{ji} = -d_{ij}$. Based on the regions R_i^+, it is not difficult to show that the discriminant function d_{ij} distinguishes the i^{th} group ordering from the j^{th} group ordering $(i, j = 1, \ldots, 6;\ i < j)$ and, thus, there are a total of 15 distinct discriminant functions that differentiate one group ordering from another.

B.2 CLASSIFYING ALL MEMBERS OF A MATCH IN LDA USING UNKNOWN MATCH EFFECT

We now discuss how the approach we develop in Appendix B.1, which focuses on matching across three groups, can be implemented using available training data consisting of $\mathbf{y}_{ik}$, the observed feature vector for the member of the k^{th} triple belonging to the i^{th} group ($i = 1, 2, 3;\ k = 1, \ldots, K$). To evaluate the classification rule in (B.2), we need to estimate the parameters $\boldsymbol{\Sigma}_*$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$. We can do so by first considering the model for the transformed stacked feature vector $\mathbf{AY}^+$ in Appendix B.1, where we showed that the models for $\mathbf{Y}^+$ and $\mathbf{AY}^+$ yielded the same linear discriminant rule. Recall from our model for $\mathbf{AY}^+$ that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\upsilon}_1 - \frac{1}{2}\boldsymbol{\upsilon}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3 = \boldsymbol{\upsilon}_1 + \frac{1}{2}\boldsymbol{\upsilon}_2$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3 = \boldsymbol{\upsilon}_2$.

We can estimate $\boldsymbol{\upsilon}_1$, $\boldsymbol{\upsilon}_2$, and $\boldsymbol{\Sigma}_*$ using the transformed training data $\begin{bmatrix} \mathbf{z}_{ik} \\ \mathbf{s}_{jk} \\ \mathbf{w}_{ik} \end{bmatrix}$, where $\mathbf{z}_{ik} = \mathbf{y}_{ik} - \frac{1}{2}(\mathbf{y}_{jk} + \mathbf{y}_{lk})$, $\mathbf{s}_{jk} = \mathbf{y}_{jk} - \mathbf{y}_{lk}$, and $\mathbf{w}_{ik} = \mathbf{y}_{ik} + \mathbf{y}_{jk} + \mathbf{y}_{lk}$ $(i, j, l = 1, 2, 3;\ i \neq j \neq l)$. Based on the facts that $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$ are unconstrained, $\boldsymbol{\Sigma}_*$ is only constrained to be positive definite, and the random vector $\mathbf{W}$ is independent of the difference vectors $\mathbf{Z}$ and $\mathbf{S}$ in our model for $\mathbf{AY}^+$, we can estimate $\boldsymbol{\upsilon}_1$, $\boldsymbol{\upsilon}_2$, and $\boldsymbol{\Sigma}_*$ solely based on the differenced training feature vector $\begin{bmatrix} \mathbf{z}_{ik} \\ \mathbf{s}_{jk} \end{bmatrix}$. Using ML estimation, the ML estimates of $\boldsymbol{\upsilon}_1$, $\boldsymbol{\upsilon}_2$, and $\boldsymbol{\Sigma}_*$ can be shown to equal $\hat{\boldsymbol{\upsilon}}_1 = \bar{\mathbf{z}}_{1\cdot} = \bar{\mathbf{y}}_{1\cdot} - \frac{1}{2}(\bar{\mathbf{y}}_{2\cdot} + \bar{\mathbf{y}}_{3\cdot})$, $\hat{\boldsymbol{\upsilon}}_2 = \bar{\mathbf{s}}_{2\cdot} = \bar{\mathbf{y}}_{2\cdot} - \bar{\mathbf{y}}_{3\cdot}$, and $\hat{\boldsymbol{\Sigma}}_* = \frac{1}{2K}\left[\frac{2}{3}\sum_{k=1}^{K}(\mathbf{z}_{1k} - \bar{\mathbf{z}}_{1\cdot})(\mathbf{z}_{1k} - \bar{\mathbf{z}}_{1\cdot})' + \frac{1}{2}\sum_{k=1}^{K}(\mathbf{s}_{2k} - \bar{\mathbf{s}}_{2\cdot})(\mathbf{s}_{2k} - \bar{\mathbf{s}}_{2\cdot})'\right]$. Hence, the ML estimates of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$ can be directly obtained.

B.3 CLASSIFYING ALL MEMBERS OF A MATCH IN LDA USING KNOWN MATCH AND COVARIATE EFFECTS

We again focus on the case where individuals are matched across three groups. Given $\mathbf{X}^+ = \mathbf{x}^+$, we assume $\mathbf{Y}^+ \sim N_{3P}(\boldsymbol{\mu}^+_{i(x)}, \boldsymbol{\Sigma}^+_{(x)})$ in the i^{th} group ordering ($i = 1, \ldots, 6$) of $\mathbf{Y}_{ind}$, $\mathbf{Y}_{sib,1}$, and $\mathbf{Y}_{sib,2}$ for a given match, where

$$\boldsymbol{\mu}^+_{1(x)} = \begin{bmatrix} \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2} \end{bmatrix},\ \boldsymbol{\mu}^+_{2(x)} = \begin{bmatrix} \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2} \end{bmatrix},\ \boldsymbol{\mu}^+_{3(x)} = \begin{bmatrix} \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2} \end{bmatrix},\ \boldsymbol{\mu}^+_{4(x)} = \begin{bmatrix} \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2} \end{bmatrix},$$

$$\boldsymbol{\mu}^+_{5(x)} = \begin{bmatrix} \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2} \end{bmatrix},\ \boldsymbol{\mu}^+_{6(x)} = \begin{bmatrix} \boldsymbol{\mu}_3+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{ind} \\ \boldsymbol{\mu}_2+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,1} \\ \boldsymbol{\mu}_1+\boldsymbol{\gamma}+\boldsymbol{\beta}\mathbf{x}_{sib,2} \end{bmatrix},\ \boldsymbol{\Sigma}^+_{(x)} = \begin{bmatrix} \boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi} & \boldsymbol{\Psi} \\ \boldsymbol{\Psi} & \boldsymbol{\Psi} & \boldsymbol{\Sigma}_{(x)}+\boldsymbol{\Psi} \end{bmatrix},$$

and $\mathbf{\Psi}$ represents the symmetric covariance matrix between any two of the three random feature vectors in that match. To ensure that $\mathbf{\Sigma}^{+}_{(x)}$ is positive definite, we assume $\mathbf{\Sigma}_{(x)}$ and $\mathbf{\Sigma}_{(x)} + 3\mathbf{\Psi}$ are positive definite.

Using the invariance property of LDA, we can equivalently consider the conditional distribution of $\mathbf{AY}^{+}$, where $\mathbf{A}$ and $\mathbf{AY}^{+} \equiv \begin{bmatrix} \mathbf{Z} \\ \mathbf{S} \\ \mathbf{W} \end{bmatrix}$ are defined as in Appendix B.1. Given $\mathbf{X}^{+} = \mathbf{x}^{+}$, $\mathbf{AY}^{+} \sim N_{3P}(\mathbf{A}\boldsymbol{\mu}^{+}_{i(x)}, \mathbf{A}\mathbf{\Sigma}^{+}_{(x)}\mathbf{A}')$ in the i^{th} group ordering $(i = 1, \ldots, 6)$, where

$$\mathbf{A}\boldsymbol{\mu}^{+}_{1(x)} = \begin{bmatrix} \boldsymbol{\upsilon}_1 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},1} \\ \boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},2} \\ \boldsymbol{\upsilon}_3 + 3\boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{\text{sum}} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}^{+}_{2(x)} = \begin{bmatrix} \boldsymbol{\upsilon}_1 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},1} \\ -\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},2} \\ \boldsymbol{\upsilon}_3 + 3\boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{\text{sum}} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}^{+}_{3(x)} = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1 + \frac{3}{4}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},1} \\ \boldsymbol{\upsilon}_1 + \frac{1}{2}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},2} \\ \boldsymbol{\upsilon}_3 + 3\boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{\text{sum}} \end{bmatrix},$$

$$\mathbf{A}\boldsymbol{\mu}^{+}_{4(x)} = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1 + \frac{3}{4}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},1} \\ -\boldsymbol{\upsilon}_1 - \frac{1}{2}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},2} \\ \boldsymbol{\upsilon}_3 + 3\boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{\text{sum}} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}^{+}_{5(x)} = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1 - \frac{3}{4}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},1} \\ \boldsymbol{\upsilon}_1 - \frac{1}{2}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},2} \\ \boldsymbol{\upsilon}_3 + 3\boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{\text{sum}} \end{bmatrix}, \quad \mathbf{A}\boldsymbol{\mu}^{+}_{6(x)} = \begin{bmatrix} -\frac{1}{2}\boldsymbol{\upsilon}_1 - \frac{3}{4}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},1} \\ -\boldsymbol{\upsilon}_1 + \frac{1}{2}\boldsymbol{\upsilon}_2 + \boldsymbol{\beta}\mathbf{x}_{\text{diff},2} \\ \boldsymbol{\upsilon}_3 + 3\boldsymbol{\gamma} + \boldsymbol{\beta}\mathbf{x}_{\text{sum}} \end{bmatrix},$$

$$\mathbf{A}\mathbf{\Sigma}^{+}_{(x)}\mathbf{A}' = \begin{bmatrix} \frac{3}{2}\mathbf{\Sigma}_{*(x)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{\Sigma}_{*(x)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 3\mathbf{\Sigma}_{1(x)} \end{bmatrix},$$

$\mathbf{x}_{\text{diff},1} = \mathbf{x}_{ind} - \frac{1}{2}(\mathbf{x}_{sib,1} + \mathbf{x}_{sib,2})$, $\mathbf{x}_{\text{diff},2} = \mathbf{x}_{sib,1} - \mathbf{x}_{sib,2}$, $\mathbf{x}_{\text{sum}} = \mathbf{x}_{ind} + \mathbf{x}_{sib,1} + \mathbf{x}_{sib,2}$, $\mathbf{\Sigma}_{*(x)} = \mathbf{\Sigma}_{(x)}$, $\mathbf{\Sigma}_{1(x)} = \mathbf{\Sigma}_{(x)} + 3\mathbf{\Psi}$, and $\boldsymbol{\upsilon}_1$, $\boldsymbol{\upsilon}_2$, and $\boldsymbol{\upsilon}_3$ are defined as in Appendix B.1.

From our assumption regarding $\mathbf{\Sigma}_{(x)}$ and $\mathbf{\Sigma}_{(x)} + 3\mathbf{\Psi}$, we have that $\mathbf{\Sigma}_{*(x)}$ and $\mathbf{\Sigma}_{1(x)}$ are both positive definite and, thus, so is $\mathbf{A}\mathbf{\Sigma}^{+}_{(x)}\mathbf{A}'$. Based on our argument in Appendix A, the fact that $\mathbf{A}\mathbf{\Sigma}^{+}_{(x)}\mathbf{A}'$ is positive definite implies that $\mathbf{\Sigma}^{+}_{(x)}$ is also positive definite. It is not difficult to show that our assumption that $\mathbf{\Psi}$ is symmetric is a sufficient condition for $\mathbf{Z}$, $\mathbf{S}$, and $\mathbf{W}$ to be mutually independent based on our conditional model for $\mathbf{AY}^{+}$.

For a given match and a given value of $\mathbf{x}^{+}$, $\mathbf{W}$ is independent from both $\mathbf{Z}$ and $\mathbf{S}$, and the conditional distribution of $\mathbf{W}$ provides no information that aids in discriminating among the six group orderings. Thus, we only need to consider the conditional distribution of $\left[\begin{smallmatrix} \mathbf{Z} \\ \mathbf{S} \end{smallmatrix}\right]$ in each group ordering when constructing our classification rule. Retaining our assumption from Appendix B.1 of equal priors for each group ordering, we can apply general covariance adjusted LDA based on the conditional distributions of $\left[\begin{smallmatrix} \mathbf{Z} \\ \mathbf{S} \end{smallmatrix}\right]$ to obtain the following classification rule:

$$R^{+}_{i(x)}: \left[\mathbf{d}^{*} - \frac{1}{2}\left(\boldsymbol{\mu}^{+(*)}_{i(x)} + \boldsymbol{\mu}^{+(*)}_{j(x)}\right)\right]' \begin{bmatrix} \frac{2}{3}\mathbf{\Sigma}^{-1}_{*(x)} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{\Sigma}^{-1}_{*(x)} \end{bmatrix} \left(\boldsymbol{\mu}^{+(*)}_{i(x)} - \boldsymbol{\mu}^{+(*)}_{j(x)}\right) > 0 \; j = 1, \ldots, 6; \; j \neq i, \tag{B.3}$$

where $\mathbf{d}^{*}$ is defined as in Appendix B.1 and $\boldsymbol{\mu}^{+(*)}_{i(x)}$ denotes the first $2P$ components of $\mathbf{A}\boldsymbol{\mu}^{+}_{i(x)}$ $(i = 1, \ldots, 6)$. A new triad with an observed feature value $\mathbf{d}^{*}$ would then be classified into the i^{th} group ordering if $\mathbf{d}^{*}$ falls into region $R^{+}_{i(x)}$ $(i = 1, \ldots, 6)$.

Based on the formulas for $\boldsymbol{\mu}_{i(x)}^{+(*)}$ and $\boldsymbol{\mu}_{j(x)}^{+(*)}$, the rule in (B.3) can be re-expressed as

$$R_{i(x)}^{+} : d_{ij(x)} > 0 \quad j = 1, \ldots, 6;\ j \neq i, \tag{B.4}$$

where

$$\begin{aligned}
d_{12(x)} &= [\mathbf{y}_{sib,1} - \mathbf{y}_{sib,2} - \boldsymbol{\beta}(\mathbf{x}_{sib,1} - \mathbf{x}_{sib,2})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3),\\
d_{13(x)} &= [\mathbf{y}_{ind} - \mathbf{y}_{sib,1} - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib,1})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),\\
d_{16(x)} &= [\mathbf{y}_{ind} - \mathbf{y}_{sib,2} - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib,2})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3),\\
d_{24(x)} &= [\mathbf{y}_{ind} - \mathbf{y}_{sib,2} - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib,2})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),\\
d_{25(x)} &= [\mathbf{y}_{ind} - \mathbf{y}_{sib,1} - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib,1})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3),\\
d_{34(x)} &= [\mathbf{y}_{sib,1} - \mathbf{y}_{sib,2} - \boldsymbol{\beta}(\mathbf{x}_{sib,1} - \mathbf{x}_{sib,2})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3),\\
d_{35(x)} &= [\mathbf{y}_{ind} - \mathbf{y}_{sib,2} - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib,2})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3),\\
d_{46(x)} &= [\mathbf{y}_{ind} - \mathbf{y}_{sib,1} - \boldsymbol{\beta}(\mathbf{x}_{ind} - \mathbf{x}_{sib,1})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3),\\
d_{56(x)} &= [\mathbf{y}_{sib,1} - \mathbf{y}_{sib,2} - \boldsymbol{\beta}(\mathbf{x}_{sib,1} - \mathbf{x}_{sib,2})]' \boldsymbol{\Sigma}_{*(x)}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),
\end{aligned}$$

$d_{14(x)} = d_{12(x)} + d_{24(x)}$, $d_{15(x)} = d_{13(x)} + d_{35(x)}$, $d_{23(x)} = d_{13(x)} - d_{12(x)}$, $d_{26(x)} = d_{24(x)} + d_{46(x)}$, $d_{36(x)} = d_{16(x)} - d_{13(x)}$, and $d_{45(x)} = d_{25(x)} - d_{24(x)}$. We can compute the other 15 $d_{ij(x)}$ functions by using the fact that $d_{ji(x)} = -d_{ij(x)}$. Using the regions $R_{i(x)}^{+}$, it can be shown that the discriminant function $d_{ij(x)}$ distinguishes the i^{th} group ordering from the j^{th} group ordering $(i, j = 1, \ldots, 6;\ i < j)$ and, thus, there are 15 distinct discriminant functions that differentiate one group ordering from another.

B.4 CLASSIFYING ALL MEMBERS OF A MATCH IN LDA USING UNKNOWN MATCH AND COVARIATE EFFECTS

With training data consisting of $(\mathbf{y}_{ik}, \mathbf{x}_{ik})$, the observed feature and covariate vectors for the member of the k^{th} triple belonging to the i^{th} group $(i = 1, 2, 3;\ k = 1, \ldots, K)$, we can estimate the parameters $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_{*(x)}$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$ that are needed to compute the classification regions in (B.4). To estimate these parameters, we consider the

conditional model for $\mathbf{Y}^+$ in Appendix B.3, as well as the conditional model for $\mathbf{A}\mathbf{Y}^+$, for which $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\upsilon}_1 - \frac{1}{2}\boldsymbol{\upsilon}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3 = \boldsymbol{\upsilon}_1 + \frac{1}{2}\boldsymbol{\upsilon}_2$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3 = \boldsymbol{\upsilon}_2$.

We first obtain a consistent estimator of $\boldsymbol{\beta}$, which we estimate from the training data based on our conditional model for $\mathbf{Y}^+$. To elaborate, let $\mathbf{Y}_{ik}$ denote the random feature vector that corresponds to the member of the k^{th} triple belonging to the i^{th} group ($i = 1, 2, 3$; $k = 1, \ldots, K$), with conditional mean $E[\mathbf{Y}_{ik}] = \boldsymbol{\mu}_i + \boldsymbol{\gamma}_k + \boldsymbol{\beta}\mathbf{x}_{ik}$. To obtain a consistent estimator of $\boldsymbol{\beta}$, we can use LS estimation to fit our assumed model for the conditional mean $E[\mathbf{Y}_{ik}]$ using the training data. The design matrix for this model is assumed to satisfy suitable conditions so that the LS estimate $\hat{\boldsymbol{\beta}}$ is unique.

Once we obtain the estimate $\hat{\boldsymbol{\beta}}$, which we view as fixed, we can estimate $\boldsymbol{\Sigma}_{*(x)}$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$ using a procedure similar to that used in Appendix B.2, where we only accounted for the effect of triple matching. Specifically, we compute the covariate adjusted differenced training feature vector $\begin{bmatrix} \mathbf{z}^*_{ik} \\ \mathbf{s}^*_{jk} \end{bmatrix}$, where $\mathbf{z}^*_{ik} = [\mathbf{y}_{ik} - \frac{1}{2}(\mathbf{y}_{jk} + \mathbf{y}_{lk})] - \hat{\boldsymbol{\beta}}[\mathbf{x}_{ik} - \frac{1}{2}(\mathbf{x}_{jk} + \mathbf{x}_{lk})]$, and $\mathbf{s}^*_{jk} = \mathbf{y}_{jk} - \mathbf{y}_{lk} - \hat{\boldsymbol{\beta}}(\mathbf{x}_{jk} - \mathbf{x}_{lk})$ ($i, j, l = 1, 2, 3$; $i \neq j \neq l$) and, based on this data, use the same ML estimation procedure as in Appendix B.2. In doing so, we obtain the ML estimates $\hat{\boldsymbol{\upsilon}}_1 = \bar{\mathbf{z}}^*_{1.} = [\bar{\mathbf{y}}_{1.} - \frac{1}{2}(\bar{\mathbf{y}}_{2.} + \bar{\mathbf{y}}_{3.})] - \hat{\boldsymbol{\beta}}[\bar{\mathbf{x}}_{1.} - \frac{1}{2}(\bar{\mathbf{x}}_{2.} + \bar{\mathbf{x}}_{3.})]$, $\hat{\boldsymbol{\upsilon}}_2 = \bar{\mathbf{s}}^*_{2.} = \bar{\mathbf{y}}_{2.} - \bar{\mathbf{y}}_{3.} - \hat{\boldsymbol{\beta}}(\bar{\mathbf{x}}_{2.} - \bar{\mathbf{x}}_{3.})$, and $\hat{\boldsymbol{\Sigma}}_{*(x)} = \frac{1}{2K}\left[\frac{2}{3}\sum_{k=1}^{K}(\mathbf{z}^*_{1k} - \bar{\mathbf{z}}^*_{1.})(\mathbf{z}^*_{1k} - \bar{\mathbf{z}}^*_{1.})' + \frac{1}{2}\sum_{k=1}^{K}(\mathbf{s}^*_{2k} - \bar{\mathbf{s}}^*_{2.})(\mathbf{s}^*_{2k} - \bar{\mathbf{s}}^*_{2.})'\right]$. From these estimates of $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$, we can obtain the corresponding estimates of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_3$, and $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_3$.

APPENDIX C

C.1 PROPERTIES OF IMPURITY MEASURE BASED GOS CRITERIA

Proposition C.1.1. *Let $\phi(p_1, \ldots, p_g)$ be a strictly concave function such that $p_i \geq 0$ $(i = 1, \ldots, g)$ and $\sum_{i=1}^{g} p_i = 1$. For $M(t) = \phi\left(P(1|t), \ldots, P(g|t)\right)$, where $P(\mathbf{Y} \in$ group $i | \mathbf{Y} \in t) = P(i|t)$,*

$$M(t) - P_L M(t_L) - P_R M(t_R) \geq 0, \tag{C.1}$$

where $P_L = P(\mathbf{Y} \in t_L | \mathbf{Y} \in t)$ and $P_R = P(\mathbf{Y} \in t_R | \mathbf{Y} \in t)$. Equality in (C.1) holds if and only if $P(i|t_L) = P(i|t_R) = P(i|t)$ $(i = 1, \ldots, g)$. If $\mathbf{Y}$ is continuous, the inequality in (C.1) is strict.

Proof. Since ϕ is strictly concave,

$$\begin{aligned} P_L M(t_L) + P_R M(t_R) &= P_L\, \phi\left(P(1|t_L), \ldots, P(g|t_L)\right) + P_R\, \phi\left(P(1|t_R), \ldots, P(g|t_R)\right) \\ &\leq \phi\left(P_L P(1|t_L) + P_R P(1|t_R), \ldots, P_L P(g|t_L) + P_R P(g|t_R)\right), \end{aligned} \tag{C.2}$$

with equality holding in (C.2) if and only if $P(i|t_L) = P(i|t_R) = P(i|t)$ $(i = 1, \ldots, g)$. Since

$$\begin{aligned} P_L &= \frac{P(\mathbf{Y} \in t_L, \mathbf{Y} \in t)}{P(\mathbf{Y} \in t)} = \frac{P(\mathbf{Y} \in t_L)}{P(\mathbf{Y} \in t)} \quad (t_L \subset t) \quad \text{and} \\ P_R &= \frac{P(\mathbf{Y} \in t_R, \mathbf{Y} \in t)}{P(\mathbf{Y} \in t)} = \frac{P(\mathbf{Y} \in t_R)}{P(\mathbf{Y} \in t)} \quad (t_R \subset t), \end{aligned}$$

it follows that

$$\begin{aligned} P_L P(i|t_L) + P_R P(i|t_R) &= \frac{\left[P(\mathbf{Y} \in \text{group } i, \mathbf{Y} \in t_L) + P(\mathbf{Y} \in \text{group } i, \mathbf{Y} \in t_R)\right]}{P(\mathbf{Y} \in t)} \\ &= \frac{P(\mathbf{Y} \in \text{group } i, \mathbf{Y} \in t)}{P(\mathbf{Y} \in t)} \\ &= P(i|t). \end{aligned} \tag{C.3}$$

From (C.3), the right hand side of the inequality in (C.2) is equal to $\phi\left(P(1|t), \ldots, P(g|t)\right)$ $= M(t)$ and, thus, $M(t) - P_L M(t_L) - P_R M(t_R) \geq 0$. Equality holds if and only if $P(i|t_L)$

$= P(i|t_R) = P(i|t)$ $(i = 1, \ldots, g)$. If $\mathbf{Y}$ is continuous, this condition will never hold and $M(t) - P_L M(t_L) - P_R M(t_R)$ will be positive. □

C.2 TREE CONSTRUCTION USING STACKED FEATURE VECTOR, ADJUSTING FOR EFFECT OF MATCHING

Based on the distribution of the stacked random feature vector $\mathbf{Y}^+_{\boldsymbol{\gamma}}$, we can also consider $\tilde{\mathbf{Y}}^+ = \begin{bmatrix} \tilde{\mathbf{Y}}_{ind} \\ \tilde{\mathbf{Y}}_{sib,1} \\ \vdots \\ \tilde{\mathbf{Y}}_{sib,g-1} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{\boldsymbol{\gamma},ind} - \boldsymbol{\gamma} \\ \mathbf{Y}_{\boldsymbol{\gamma},sib,1} - \boldsymbol{\gamma} \\ \vdots \\ \mathbf{Y}_{\boldsymbol{\gamma},sib,g-1} - \boldsymbol{\gamma} \end{bmatrix}$, which has been adjusted for the effect of matching and has known CDF $F^{(i_1)}_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{c}}_{ind}) \times F^{(i_2)}_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{c}}_{sib,1}) \times \ldots F^{(i_g)}_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{c}}_{sib,g-1}) \equiv F^{(l)}_{\tilde{\mathbf{Y}}^+}(\tilde{\mathbf{c}}_{ind}, \tilde{\mathbf{c}}_{sib,1}, \ldots, \tilde{\mathbf{c}}_{sib,g-1})$ in the l^{th} group ordering $(i_1, i_2, \ldots, i_g = 1, \ldots, g;\ i_1 \neq i_2 \neq \cdots \neq i_g;\ l = 1, \ldots, g!)$.

As was previously stated in Section 4.4.1, it is equally likely that each member of a given match belongs to either of the two groups and the labeling of a member as an individual or as any of the $g - 1$ siblings is assumed to be completely random. Therefore, we assume that each group ordering is equally likely, i.e., the prior probability of each group ordering is equal to $1/g!$. Based on this assumption and our assumption of equal misclassification costs, we can construct the adjusted tree $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ using the traditional population-based approach in Section 4.1.2.1 by replacing the probabilities $P^{(i)}(\mathbf{Y} \in t)$, $P^{(i)}(\mathbf{Y} \in t_L)$, and $P^{(i)}(\mathbf{Y} \in t_R)$ used to construct T' with the probabilities $P^{(l)}(\tilde{\mathbf{Y}}^+ \in t)$, $P^{(l)}(\tilde{\mathbf{Y}}^+ \in t_L)$, and $P^{(l)}(\tilde{\mathbf{Y}}^+ \in t_R)$ based on the CDF of $\tilde{\mathbf{Y}}^+$ in the l^{th} group ordering, where $P^{(l)}(\tilde{\mathbf{Y}}^+ \in t) = P(\tilde{\mathbf{Y}}^+ \in t | \tilde{\mathbf{Y}}^+ \in \text{ordering } l)$. We can use the following rule to assign each terminal node t of $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ to the l^{th} group ordering of $\tilde{\mathbf{Y}}^+$:

$$R^+_l : \left\{ t : P^{(l)}(\tilde{\mathbf{Y}}^+ \in t) > P^{(j)}(\tilde{\mathbf{Y}}^+ \in t) \right\}, \quad j = 1, \ldots, g!;\ j \neq l. \tag{C.4}$$

If the observed adjusted feature data for a new match, $\tilde{\mathbf{y}}^+$, falls into a terminal node of $T'^{\text{adj}(\boldsymbol{\gamma}^+)}$ that has been assigned to the l^{th} group ordering according to the rule in (C.4), then we simultaneously classify all members in that match into the l^{th} group ordering.

APPENDIX D

APPLICATION OF DIFFERENCING AND STACKED LDA APPROACHES TO KONOPASKE DATA

To show the difference between the differencing and stacked approaches in Sections 3.5.2.2 and 3.5.2.3, we applied these two approaches to the following three biomarkers measured in the Konopaske et al. brain tissue study: oligodendrocyte number, oligodendrocyte density, and the ratio of oligodendrocyte number to glial cell number (oligodendrocyte ratio). The linear discriminant functions obtained from the differencing approach are displayed in Table D1.

Table D1: Linear Discriminant Functions for Haloperidol (H) vs. Olanzapine (O), Haloperidol vs. Sham (S), and Olanzapine vs. Sham (Differencing Approach)

Biomarker	Coefficient (H vs. O)	Coefficient (H vs. S)	Coefficient (O vs. S)
Oligodendrocyte Number	0.00000011	-0.00000047	-0.00000059
Oligodendrocyte Density	-0.0008469	0.0011501	0.001997
Oligodendrocyte Ratio	36.67692	2.71122	-33.9657

In our application of the stacked approach, we have the following six treatment group orderings:

Member	Ordering 1	Ordering 2	Ordering 3	Ordering 4	Ordering 5	Ordering 6
Individual	haloperidol	haloperidol	olanzapine	olanzapine	sham	sham
Sibling 1	olanzapine	sham	haloperidol	sham	haloperidol	olanzapine
Sibling 2	sham	olanzapine	sham	haloperidol	olanzapine	haloperidol

The estimates of the linear discriminant functions d_{ij} $(i, j = 1, \ldots, 6;\ i < j)$ in (B.2) are provided in Tables D2 through D4.

Table D2: Linear Discriminant Functions for Konopaske Data (Stacked Approach)
(Oligo Number - Oligodendrocyte Number, Oligo Density - Oligodendrocyte Density,
Oligo Ratio - Oligodendrocyte Ratio)

Member	Biomarker	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}
Individual	Oligo Number	0	0.00000011	0.00000011	-0.00000047	-0.00000047
	Oligo Density	0	-0.0008469	-0.0008469	0.0011501	0.0011501
	Oligo Ratio	0	36.67692	36.67692	2.71122	2.71122
Sibling 1	Oligo Number	-0.00000059	-0.00000011	-0.00000059	-0.00000011	0
	Oligo Density	0.001994	0.0008469	0.001994	0.0008469	0
	Oligo Ratio	-33.9657	-36.67692	-33.9657	-36.67692	0
Sibling 2	Oligo Number	0.00000059	0	0.00000047	0.00000059	0.00000047
	Oligo Density	-0.001994	0	-0.0011471	-0.001997	-0.0011501
	Oligo Ratio	33.9657	0	-2.71122	33.9657	-2.71122

Table D3: Linear Discriminant Functions for Konopaske Data cont. (Stacked Approach)

Member	Biomarker	d_{23}	d_{24}	d_{25}	d_{26}	d_{34}
Individual	Oligo Number	0.00000011	0.00000011	-0.00000047	-0.00000047	0
	Oligo Density	-0.0008469	-0.0008469	0.0011501	0.0011501	0
	Oligo Ratio	36.67692	36.67692	2.71122	2.71122	0
Sibling 1	Oligo Number	0.00000047	0	0.00000047	0.00000059	-0.00000047
	Oligo Density	-0.0011471	0	-0.0011501	-0.001997	0.0011472
	Oligo Ratio	-2.71122	0	-2.71122	33.9657	2.71122
Sibling 2	Oligo Number	-0.00000059	-0.00000011	0	-0.00000011	0.00000047
	Oligo Density	0.001994	0.0008469	0	0.0008469	-0.0011472
	Oligo Ratio	-33.9657	-36.67692	0	-36.67692	-2.71122

Table D4: Linear Discriminant Functions for Konopaske Data cont. (Stacked Approach)

Member	Biomarker	d_{35}	d_{36}	d_{45}	d_{46}	d_{56}
Individual	Oligo Number	-0.00000059	-0.00000059	-0.00000059	-0.00000059	0
	Oligo Density	0.001997	0.001997	0.001997	0.001997	0
	Oligo Ratio	-33.9657	-33.9657	-33.9657	-33.9657	0
Sibling 1	Oligo Number	0	0.00000011	0.00000047	0.00000059	0.00000011
	Oligo Density	0	-0.0008469	-0.0011501	-0.001997	-0.000847
	Oligo Ratio	0	36.67692	-2.71122	33.9657	36.67692
Sibling 2	Oligo Number	0.00000059	0.00000047	0.00000011	0	-0.00000011
	Oligo Density	-0.001997	-0.0011501	-0.0008469	0	0.000847
	Oligo Ratio	33.9657	-2.71122	36.67692	0	-36.67692

We can compute the other 15 discriminant functions d_{ij} $(i, j = 1, \ldots, 6;\ i > j)$ by using the fact that $d_{ji} = -d_{ij}$. In addition, using the classification regions obtained from the differencing and stacked approaches, we have that the cross validated correct classification rates obtained from these two approaches are 44% and 50%, respectively.

The three discriminant functions in Table D1 can be used, after appropriate standardization, to identify which of the three biomarkers under consideration best discriminate among the haloperidol, olanzapine, and sham treatment groups. On the other hand, although we have that the discriminant functions d_{ij} in Tables D2 to D4 $(i, j = 1, \ldots, 6;\ i < j)$ discriminate between the i^{th} and j^{th} treatment group orderings, it is not readily apparent that we can use this kind of information to determine which of the three biomarkers best distinguish among the three treatment groups.

In examining the discriminant functions and correct classification rates obtained from applying our differencing and stacked approaches to the Konopaske et al. biomarker data, we clearly see that as was the case in the population setting, these two approaches yield completely different types of results in practice from both a discrimination and classification standpoint.

BIBLIOGRAPHY

[1] Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis (2nd edition)*. New York, NY: John Wiley & Sons.

[2] Ben-Haim, Z. and Dvorkind, T. (2004). Majorization and applications to optimization. Technical report, CiteSeerX - Scientific Literature Digital Library and Search Engine, United States. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.4.9813.

[3] Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 24:123–140.

[4] Breiman, L. (1996b). Bias, variance, and arcing classifiers. Technical Report 460, University of California, Berkeley, CA 94720.

[5] Breiman, L. (1996c). Technical note: Some properties of splitting criteria. *Machine Learning*, 24:41–47.

[6] Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

[7] Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth Int. Group.

[8] Cochran, W. and Bliss, C. (1948). Discriminant functions with covariance. *Annals of Mathematical Statistics*, 19:151–176.

[9] Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

[10] Friedman, J. (1977). A recursive partitioning decision rule for nonparametric classifcation. *IEEE Trans Computers*, C-26:404–408.

[11] Friedman, J. (1989). Regularized discriminant analysis. *Journal of the Americal Statistical Association*, 84:165–175.

[12] Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328.

[13] Hand, D. (1997). *Construction and Assessment of Classification Rules*. Chichester, UK: John Wiley & Sons.

[14] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. New York, NY: Springer Verlag.

[15] Knable, M., Barci, B., Bartko, J., Webster, M., and Torrey, E. (2002). Molecular abnormalities in the major psychiatric illnesses: classification and regression (crt) analysis of post-mortem prefrontal markers. *Molecular Psychiatry*, 7:392–404.

[16] Knable, M., Torrey, E., Webster, M., and Bartko, J. (2001). Multivariate analysis of prefrontal cortical data from the stanley foundation neuropathology consortium. *Brain Research Bulletin*, 55(5):651–659.

[17] Konopaske, G., Dorph-Petersen, K., Sweet, R., Pierri, J., Zhang, W., Sampson, A., and Lewis, D. (2008). Effect of chronic antipsychotic exposure on astrocyte and oligodendrocyte numbers in macaque monkeys. *Biological Psychiatry*, 63(8):759–765.

[18] Koutkova, H. (1992). On estimable and locally estimable functions in the non-linear regression model. *Kybernetika*, 28(2):120–128.

[19] Lachenbruch, P. (1977). Covariance adjusted discriminant functions. *Annals of the Institute of Statistical Mathematics*, 29:247–257.

[20] Li, Q. and Racine, J. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26:423–434.

[21] Marshall, A. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications.*, volume 143 of *Mathematics in Science and Engineering Series*. New York, NY: Academic Press.

[22] McLachlan, G. (1976). The bias of the apparent error rate in discriminant analysis. *Biometrika*, 63(2):239–244.

[23] McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York, NY: John Wiley & Sons.

[24] Morgan, J. and Messenger, R. (1973). THAID: a sequential search program for the analysis of nominal scale dependent variables. Technical report, Institute for Social Research, University of Michigan, Ann Arbor.

[25] Morgan, J. and Sonquist, J. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302):415–434.

[26] Mueller, R. and Cozad, J. (1988). Standardized discriminant coefficients: which variance estimate is appropriate? *Journal of Educational Statistics*, 13(4):313–318.

[27] Mueller, R. and Cozad, J. (1993). Standardized discriminant coefficients: a rejoinder. *Journal of Educational Statistics*, 18(1):108–114.

[28] Peracchi, F. (2002). On estimating conditional quantiles and distribution functions. *Computational Statistics & Data Analysis*, 38(4):433–447.

[29] Rawlings, R., Graubard, B., Teper, S., Ryback, R., and Eckardt, M. (1986). Conditional quadratic discrimination in the identification of biological markers for disease screening. *Biometrics*, 28(8):957–964.

[30] Rokach, L. and Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications*, volume 69 of *Series in Machine Perception and Artificial Intelligence*. Singapore: World Scientific.

[31] Shang, N. and Breiman, L. (1996). Distribution based trees are more accurate. In *Proc. of the Int. Conf. on Neural Information Processing*, volume 1, pages 133–138.

[32] Shih, Y. (1999). Families of splitting criteria for classification trees. *Statistics and Computing*, 9:309–315.

[33] Sweet, R., Bergen, S., Sun, Z., Marcsisin, M., Sampson, A., and Lewis, D. (2007). Anatomical evidence of impaired feedforward auditory processing in schizophrenia. *Biol Psychiatry*, 61:854–864.

[34] Sweet, R., Bergen, S., Sun, Z., Sampson, A., Pierri, J., and Lewis, D. (2004). Pyramidal cell size reduction in schizophrenia: evidence for involvement of auditory feedforward circuits. *Biol Psychiatry*, 55:1128–1137.

[35] Sweet, R., Henteleff, R., Zhang, W., Sampson, A., and Lewis, D. (2008). Reduced dendritic spine density in auditory cortex of subjects with schizophrenia. *Neuropsychopharmacology*, 34:374–389.

[36] Sweet, R., Pierri, J., Auh, S., Sampson, A., and Lewis, D. (2003). Reduced pyramidal cell somal volume in auditory association cortex of subjects with schizophrenia. *Neuropsychopharmacology*, 28:599–609.

[37] Tu, X., Kowalski, J., Randall, J., Mendoza-Blanco, J., Shear, M., Monk, T., Frank, E., and Kupfer, D. (1997). Generalized covariance-adjusted discriminants: perspective and application. *Biometrics*, 53:900–909.

[38] Wang, D., Klatzky, R., Wu, B., Weller, G., Sampson, A., and Stetten, G. (2009). Fully automated common carotid artery and internal jugular vein identification and tracking using b-mode ultrasound. *IEEE Trans Biomed Eng*, 56(6):1691–1699.

[39] Wang, X., Lin, Y., Song, C., Sibille, E., and Tseng, G. A statistical framework to integrate weak-signal microarray studies adjusted for confounding variables with application to major depressive disorder. Manuscript submitted for publication.

[40] Wu, Q. (2007). *Clustering methodologies with applications to integrative analyses of postmortem tissue studies in schizophrenia*. PhD thesis, University of Pittsburgh, Pittsburgh, PA, USA. http://etd.library.pitt.edu/ETD/available/etd-08062007-164618/.