Biostatistics (2018) **19**, 1, *pp.* 42–53
doi:10.1093/biostatistics/kxx015
Advance Access publication on May 17, 2017

Covariate adjusted classification trees

JOSEPHINE K. ASAFU-ADJEI*

Department of Biostatistics, School of Nursing, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
jasafuad@email.unc.edu

ALLAN R. SAMPSON

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA

SUMMARY

In studies that compare several diagnostic groups, subjects can be measured on certain features and classification trees can be used to identify which of them best characterize the differences among groups. However, subjects may also be measured on additional covariates whose ability to characterize group differences is not meaningful or of interest, but may still have an impact on the examined features. Therefore, it is important to adjust for the effects of covariates on these features. We present a new semi-parametric approach to adjust for covariate effects when constructing classification trees based on the features of interest that is readily implementable. An application is given for postmortem brain tissue data to compare the neurobiological characteristics of subjects with schizophrenia to those of normal controls. We also evaluate the performance of our approach using a simulation study.

Keywords: Classification trees; Covariates; Features; Postmortem tissue studies; Schizophrenia.

1. INTRODUCTION

One may want to examine a set of characteristics or features measured on subjects belonging to one of multiple groups and attempt to determine a subset of these features that is superior in characterizing the differences among the examined groups. However, the ability of these features to do so may be obscured by the effects of covariates that are also measured, but are not of interest with respect to characterizing the group differences. In this case, this subset of features may be more accurately identified if the confounding effects of these covariates can be somehow adjusted for or, in other words, removed (Tu *and others*, 1997).

We were motivated to explore approaches for adjusting for covariate effects in the context of discrimination techniques due to our analyses of human postmortem brain tissue studies conducted in the Conte Center for the Neuroscience of Mental Disorders (CCNMD) in the Department of Psychiatry at the University of Pittsburgh. This center focuses on understanding the neurobiology of schizophrenia. In a typical CCNMD study, schizophrenia subjects and normal controls are measured not only on different brain biomarkers of scientific interest but also on covariates including subject's age at death, gender, brain tissue storage time (the amount of time brain tissue has been held in storage), and postmortem interval (PMI), that is, elapsed time between time of death and time of tissue collection. The primary aim in these

*To whom correspondence should be addressed.

studies is to identify those brain biomarkers that best characterize the differences between the schizophrenia and control diagnostic groups, without the effects of covariates not intrinsically meaningful to our CCNMD collaborators in understanding schizophrenia. In this setting, removing the confounding effects of these covariates on the examined brain biomarkers creates a clearer picture of which biomarkers truly characterize the differences between schizophrenia subjects and controls.

There is an extensive literature dealing with discrimination methodology. Parametric approaches include logistic discriminant analysis (Tu *and others*, 1997), as well as linear discriminant analysis (LDA), which is based on the assumption of joint normality among features. More recently, Li *and others* (2003) formulated extensions of LDA to nonlinear discriminant analysis using kernel function operators. One nonparametric discrimination approach is that of constructing classification trees, which was initially developed by various authors including Breiman, Friedman, Olshen, and Stone (Breiman *and others*, 1984) (hereafter denoted as BFOS). Recent extensions of classification trees include the approach developed by Kim and Loh (2003) which incorporates LDA and statistical testing methods, and the tree ensemble approach of Random Forests developed by Breiman (2001a). Other nonparametric discrimination approaches include support vector machines (Cortes and Vapnik, 1995).

There are currently several approaches to adjust for covariate effects when assessing the association between a set of feature variables and a response variable, for example, a classification variable. Crainiceanu *and others* (2008), Wang *and others* (2012), and Wilson and Reich (2014) developed methods for identifying optimal sets of covariates to adjust for. Friston *and others* (1995) developed an adjustment approach for statistical parametric maps of neuroimaging data that consists of fitting a linear model to each outcome while controlling for covariate effects. Based on the model, residuals are computed and the brain map is based on these residuals. Nonyane and Foulkes (2008) used a similar approach for Random Forests. Yet, the available literature that clearly develops the theoretical framework of adjusting for covariate effects on feature variables in the context of discrimination approaches appears to be limited to discriminant analysis. Specifically, early methods to adjust for covariate effects on feature data were developed for LDA by Cochran and Bliss (1948), with further developments by Lachenbruch (1977) and Tu *and others* (1997) (hereafter, denoted as L&T), as well as Asafu-Adjei (2011); Asafu-Adjei *and others* (2013). The core of these methodologies is to adjust for covariate effects using the conditional distributions of the feature data for a given covariate value in a population setting. The computational implementation of these adjustment methodologies is equivalent to fitting a linear or nonlinear model to the training feature data controlling for group and covariate effects, computing the resulting residual feature values in each group, and then basing the discriminant analysis on these residuals. Tu *and others* (1997) and Dukart *and others* (2011) also implemented this covariate adjustment approach for logistic discriminant analysis and support vector machines, respectively, but neither of these approaches provide details for their theoretical basis.

Therefore, there does not appear to be any such covariate adjustment method that is available and theoretically justified for classification trees, whose advantages over other discrimination methods include its simplicity and ease of interpretation. As a result, one finds papers such as Knable *and others* (2002) that use classification trees to identify post-mortem brain tissue biomarkers that best characterize the differences among three mental disorder diagnostic groups and controls, but fail to account for the effect that these measured covariates can have on the biomarker measurements. To lay the groundwork for our proposed modification to classification trees, we first summarize the BFOS construction method from a more general, population-based perspective, amplifying the initial approaches taken by Shang and Breiman (1996) in their development of distribution based trees. We do this primarily to provide a theoretical basis for incorporating the covariates into the BFOS construction approach using the conditional distributions of the feature data.

There are intricacies, as well as computational issues, in this implementation that soon become apparent. One such intricacy is whether or not to allow the subset of features identified as best characterizing the

group differences to depend on the covariates. Later in our discussion, we show that this can happen when covariates have differential effects on the examined features in the different groups. For example, in pre-processing the data in a post-mortem tissue study, there may be situations where subject's age at death affects the biomarkers differentially by diagnostic group, so that it makes sense for the biomarkers identified as best characterizing the diagnostic group differences to depend on the subject's age at their time of death. However, in the neurobiological studies we have collaborated on, interest has mainly been expressed in identifying a set of such biomarkers that remains the same regardless of subject's age at death or any other covariate value.

Therefore, we focus on developing a model for the conditional distribution of the feature data that allows us to adjust for covariate effects when constructing a tree based on the feature data, while yielding a subset of features that best characterizes the differences among groups and does not depend on covariate values. Specifically, in the spirit of L&T, we formulate a semiparametric model that allows us to fit a linear or nonlinear model to the training feature data that controls for group and covariate effects. Based on this model, we compute the residual (or covariate adjusted) feature values in each group, and construct a covariate adjusted classification tree (COVACT) by analyzing these residual values using standard classification tree software. Although this approach may make sense intuitively, incorporating the conditional distribution of the feature data ensures that considering these residual values in COVACT is theoretically valid and not simply an ad-hoc approach.

In Section 2, we begin by briefly discussing the motivation behind examining the conditional distribution of the feature data in our proposed approach and by giving some of the notation used throughout our discussion. We then provide our population-based formulation of the BFOS construction approach in Section 3, where we also develop a tree construction approach for the case where the feature data come from one of g known distributions conditional on a given covariate value. In Section 4, we detail our proposed COVACT construction approach. We apply COVACT to postmortem brain tissue data in Section 5, and use a simulation study to evaluate its performance in Section 6. In Section 7, we conclude with a final discussion.

2. MOTIVATION AND NOTATION

2.1. *The effects of covariates in discrimination*

Our primary goal is to accurately identify which features best characterize the differences among groups, without the masking effects of covariates that are not of interest with respect to distinguishing among groups. Examining the conditional distribution of the feature data for a given covariate value allows for the estimation and adjustment of these covariate effects. Demonstrations of this are depicted in Figure 1a of Tu *and others* (1997) and in supplementary material available at *Biostatistics* online, which illustrate that the true ability of a feature to distinguish among groups is much more apparent if one examines that feature's values adjusted to a common covariate value.

2.2. *Notation*

Let $\mathbf{Y} = (Y_1, \ldots, Y_P)$ denote a random real-valued feature vector corresponding to a set of possible measurement values for an observed subject whose group membership is unknown. We then define G as a random variable denoting group membership ($G = 1, \ldots, g$) for this subject and π_i as the prior probability of membership in group i (i.e. $\pi_i = P(G = i)$). In addition, we let $\mathbf{Y}_i$ denote $\mathbf{Y}$ given membership in group i, so that $\mathbf{Y}_i$ has a distinct population cumulative distribution function (CDF) $F_{\mathbf{Y}|i}(\cdot)$ in the i^{th} group. In the context of decision trees, the entire feature space, which contains all possible values for $\mathbf{Y}$, is defined as the root node t_0 of a particular tree and all distinct subsets of this feature space $\{t_1, t_2, \ldots\}$ as descendant nodes.

For a particular node t, that is, subset of the feature space, let $P(\mathbf{Y}_i \in t)$ denote $P(\mathbf{Y} \in t|G = i)$, which we can compute using $F_{\mathbf{Y}|i}(\cdot)$. We then have that

$$P(G = i|\mathbf{Y} \in t) = \pi_i P(\mathbf{Y}_i \in t)/P(\mathbf{Y} \in t) \tag{2.1}$$

and

$$P(\mathbf{Y} \in t) = \sum_{i=1}^{g} \pi_i P(\mathbf{Y}_i \in t). \tag{2.2}$$

Suppose we consider a set of random covariates $\mathbf{X} = (X_1, \ldots, X_S)$ that is related to the examined features. Let $\mathbf{Y}_{\mathbf{x},i}$ denote $\mathbf{Y}_i$ conditional on $\mathbf{X} = \mathbf{x}$, with population CDF $F_{\mathbf{Y}|\mathbf{x},i}(\cdot)$. In this context, we assume that $\mathbf{X}$ has no relevance in characterizing differences among groups and, thus, specify the group priors π_i to be independent of $\mathbf{x}$. For a given $\mathbf{x}$, we denote $P(\mathbf{Y} \in t|\mathbf{X} = \mathbf{x})$ as $P(\mathbf{Y}_\mathbf{x} \in t)$, $P(\mathbf{Y} \in t|G = i, \mathbf{X} = \mathbf{x})$ as $P(\mathbf{Y}_{\mathbf{x},i} \in t)$, and $P(G = i|\mathbf{Y} \in t, \mathbf{X} = \mathbf{x})$ as $P_\mathbf{x}(G = i|\mathbf{Y} \in t)$, so that

$$P_\mathbf{x}(G = i|\mathbf{Y} \in t) = \frac{P_\mathbf{x}(G = i, \mathbf{Y} \in t)}{P(\mathbf{Y}_\mathbf{x} \in t)} = \frac{\pi_i P(\mathbf{Y}_{\mathbf{x},i} \in t)}{P(\mathbf{Y}_\mathbf{x} \in t)} \tag{2.3}$$

and

$$P(\mathbf{Y}_\mathbf{x} \in t) = \sum_{i=1}^{g} P_\mathbf{x}(G = i, \mathbf{Y} \in t) = \sum_{i=1}^{g} \pi_i P(\mathbf{Y}_{\mathbf{x},i} \in t). \tag{2.4}$$

3. Classification trees: A population-based formulation

Most presentations of classification trees in the literature do not treat it from a population-based viewpoint, with the exception of a few approaches such as those taken by Shang and Breiman (1996). For instance, BFOS focus on empirical approaches towards tree construction and do not present the underlying population distribution framework. In Section 3.1, we provide a brief unified summary of the BFOS approach to classification tree construction purely from a population perspective. This presentation allows us to incorporate covariate effects into our tree construction in Section 3.2 by working with the conditional distributions of the feature data.

3.1. *Traditional procedure*

For convenience, we assume unit misclassification costs throughout. To initiate construction of a classification tree, a feature element Y_k and a cut point $c \in \mathbb{R}$ are chosen to split the feature space into two complementary descendant nodes t_L and t_R, such that a goodness of split (GOS) criterion defined by the split $Y \leq c$ is maximized over all feature elements and values of c. This splitting procedure is then applied recursively to t_L, t_R, and all subsequent descendant nodes until further splitting ceases to substantially increase group purity, as defined by some user-defined criterion. Once splitting stops, let T' denote the resulting tree consisting of nodes split using the chosen GOS criterion and nodes that are not split, which are termed as terminal nodes. For instance, if the root node t_0 is split while its descendant nodes t_L and t_R are not, the final tree T' would consist of the root node t_0 and the terminal nodes t_L and t_R. If $\mathbf{Y}$ for a particular subject falls into terminal node t, then that subject is assigned to group i if $P(G = i|\mathbf{Y} \in t) > P(G = j|\mathbf{Y} \in t)$ or equivalently, if $\pi_i P(\mathbf{Y}_i \in t) > \pi_j P(\mathbf{Y}_j \in t)$ $(j = 1, \ldots, g; j \neq i)$.

In choosing the optimal split for each node, one GOS criterion that has been used is the *two-ing* criterion (defined for empirical data by BFOS and easily extendable for the population case). Another GOS criterion is based on a measure of impurity for node t, denoted $M(t)$, where $M(t) = \phi\,(P(G = 1|\mathbf{Y} \in t), \ldots, P(G = g|\mathbf{Y} \in t))$ for a specified impurity function $\phi(.)$, which is typically chosen to be strictly concave on the unit simplex (see BFOS; Asafu-Adjei, 2011, for more details). In the case of population distributions, we rephrase this latter GOS criterion as

$$M(t) - \{P(\mathbf{Y} \in t_L|\mathbf{Y} \in t)M(t_L) + P(\mathbf{Y} \in t_R|\mathbf{Y} \in t)M(t_R)\} \tag{3.1}$$

(see BFOS). Intuitively, the bracketed quantity in (3.1) measures the within group heterogeneity of the left and right descendant nodes of node t. We want this quantity to be no larger than the measure $M(t)$ of within group heterogeneity of node t. This way, the descendant nodes of t become more internally homogeneous, or at least no more heterogeneous, with respect to group compared with node t. Examples of impurity measures $M(t)$ include the Gini and Deviance indices, two functions of $P(G = i|\mathbf{Y} \in t)$ that are strictly concave (see BFOS), which ensures that the impurity of node t is always decreased when it is split for continuous $\mathbf{Y}$ (proof provided in supplementary material available at *Biostatistics* online).

It is worth noting that the formulation of the BFOS approach from a population-based perspective allows for a parametric alternative to the standard non-parametric BFOS approach. One can assume that $\mathbf{Y}$ comes from some parametric, for example, normal, distribution and use the training data to estimate any unknown parameters (see Asafu-Adjei, 2011, for an illustration). Population classification trees also have a very useful monotone invariance property (see supplementary material available at *Biostatistics* online for more details). This property states that transforming each feature using separate strictly monotone functions does not affect the tree structure, that is, the order of splitting features and direction of splits for a particular tree.

With available data, the probabilities required for this construction procedure can be estimated non-parametrically by computing the corresponding empirical CDFs. This is actually the standard approach to the BFOS tree construction method and can be readily implemented using various software packages. To account for the fact that trees initially constructed using this non-parametric approach may substantially overfit the training data, BFOS developed a minimal cost complexity pruning approach to remove specific nodes, resulting in trees with minimized misclassification rates (see BFOS for more details). In other words, the use of minimal cost complexity pruning, which we note does not involve any formal statistical testing, results in trees with features (and their corresponding splits) that are most useful in predicting group membership and, thus, characterizing group differences.

3.2. *Conditional classification tree*

Based on our formulation in Section 3.1, we construct a tree that conditions on $\mathbf{X}$ in a straightforward manner. To obtain a conditional tree $T'^{(\mathbf{x})}$ for covariate value $\mathbf{x}$, we replace the marginal probabilities $P(\mathbf{Y} \in t)$ and $P(G = i|\mathbf{Y} \in t)$ with their conditional counterparts $P(\mathbf{Y}_\mathbf{x} \in t)$ and $P_\mathbf{x}(G = i|\mathbf{Y} \in t)$ defined in Section 2.2. Note that the relations in (2.3) and (2.4) are identical to those in (2.1) and (2.2), respectively, where $P(\mathbf{Y}_i \in t)$ is replaced with $P(\mathbf{Y}_{\mathbf{x},i} \in t)$. Thus, assuming that $F_{\mathbf{Y}|\mathbf{x},i}(.)$ and π_i are known for a given $\mathbf{x}$, we can construct $T'^{(\mathbf{x})}$ in the same manner used to construct T' in Section 3.1 by simply replacing $P(\mathbf{Y}_i \in t)$ with $P(\mathbf{Y}_{\mathbf{x},i} \in t)$.

3.3. *Conditional procedure issues*

Although this construction approach accounts for the relationship between the examined features and covariates that are assumed to have no importance in characterizing group differences (i.e., π_i not depending

on $\mathbf{x}$), it is still possible that the resulting tree structure can depend on $\mathbf{x}$. As noted in Section 1, there may be settings where one expects the tree structure to depend on covariate values, e.g., where the covariate effect on the examined features is differentially expressed across groups. However, in our noted neurobiological settings, we do not want the neurobiological features identified as best characterizing the differences among groups to depend on any particular covariate value $\mathbf{x}$ (even though the cut points may). In Section 4, we introduce a specific population model that accounts for covariate effects on the examined features and allows for the development of a tree structure that is independent of $\mathbf{x}$.

4. COVACT

4.1. *Linear invariance property*

The following invariance property is crucial in our model development (proof provided in supplementary material available at *Biostatistics* online).

Result 1: Let $\mathbf{Y}_i$ and $\mathbf{Y}_{\mathbf{x},i}$ have CDFs $F_{\mathbf{Y}|i}(.)$ and $F_{\mathbf{Y}|\mathbf{x},i}(.)$, respectively, in the i^{th} group. Suppose that $\mathbf{Y}_{\mathbf{x},i}$ is equal in distribution to $\mathbf{Y}_i + \boldsymbol{\xi}(\mathbf{x})$ (i.e., $\mathbf{Y}_{\mathbf{x},i} \overset{d}{=} \mathbf{Y}_i + \boldsymbol{\xi}(\mathbf{x})$), where $\boldsymbol{\xi}(\mathbf{x}) = (\xi_1(\mathbf{x}), \ldots, \xi_P(\mathbf{x}))'$ is a known function of $\mathbf{x}$ that does not depend on group. For a chosen GOS criterion, let $T'^{(\mathbf{x}_a)}$ be the classification tree based on $F_{\mathbf{Y}|\mathbf{x}_a,1}(.), \ldots, F_{\mathbf{Y}|\mathbf{x}_a,g}(.)$ for covariate value $\mathbf{x}_a$ and $T'^{(\mathbf{x}_b)}$ be the classification tree based on $F_{\mathbf{Y}|\mathbf{x}_b,1}(.), \ldots, F_{\mathbf{Y}|\mathbf{x}_b,g}(.)$ 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 cut points 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 $\mathbf{c}_{T'(\mathbf{x}_b)} = \mathbf{c}_{T'(\mathbf{x}_a)} - \boldsymbol{\xi}(\mathbf{x}_a) + \boldsymbol{\xi}(\mathbf{x}_b)$.

Therefore, if the conditional distribution of $\mathbf{Y}_{\mathbf{x},i}$ is a location shift of the distribution of $\mathbf{Y}_i$ by the known value $\boldsymbol{\xi}(\mathbf{x})$, then the set of splitting variables for the conditional tree $T'^{(\mathbf{x})}$ does not depend on $\mathbf{x}$. One may also think of the linear invariance property as stating that if the effects of covariates on the feature data do not depend on group, then there is no interaction between features and covariates for predicting group membership based on our tree. Verifying the assumption that the effects of covariates on the examined features are not group dependent can be done by verifying that there are no significant group by covariate interactions for any of the features.

4.2. *Population case*

For the remainder, we assume that $\mathbf{Y}_{\mathbf{x},i}$ has CDF $F_{\mathbf{Y}|\mathbf{x},i}(\mathbf{c}) = F_{\mathbf{Y}|i}(\mathbf{c} - \boldsymbol{\xi}(\mathbf{x};\Theta))$ for a given $\mathbf{x}$, where the function $\boldsymbol{\xi}(\mathbf{x};\Theta) = (\xi_1(\mathbf{x}|\boldsymbol{\theta}_1), \ldots, \xi_P(\mathbf{x}|\boldsymbol{\theta}_P))'$ is a known smooth function of $\mathbf{x}$ and parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ that is independent of group. In the spirit of L&T, we assume that $\boldsymbol{\xi}(\mathbf{x};\Theta)$ is a parametric function of $\mathbf{x}$ so that we can use $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ to estimate the magnitude of each covariate effect on the feature data. Based on our model for $F_{\mathbf{Y}|\mathbf{x},i}(\mathbf{c})$, we have that if $\mathbf{Y}_i$ has CDF $F_{\mathbf{Y}|i}(.)$, then $\mathbf{Y}_{\mathbf{x},i} \overset{d}{=} \mathbf{Y}_i + \boldsymbol{\xi}(\mathbf{x};\Theta)$. As in Section 3.1, we also assume unit misclassification costs.

Alternatively, we may examine the distribution of the covariate adjusted feature vector $\tilde{\mathbf{Y}}_i = \mathbf{Y}_{\mathbf{x},i} - \boldsymbol{\xi}(\mathbf{x};\Theta)$ in the i^{th} group, so that $\tilde{\mathbf{Y}}_i$ has CDF $F_{\tilde{\mathbf{Y}}|i}(.) \equiv F_{\mathbf{Y}|i}(.)$. Clearly, the conditional probabilities $P(\mathbf{Y}_{\mathbf{x},i} \in t)$ obtained from $F_{\mathbf{Y}|\mathbf{x},i}(.)$ can equivalently be expressed as $P(\tilde{\mathbf{Y}}_i \in t)$ obtained from $F_{\tilde{\mathbf{Y}}|i}(.)$. Once we compute these probabilities, we can construct the covariate adjusted tree $T'^{\text{adj}(\mathbf{x})}$ using the traditional approach in Section 3.1 by simply replacing $P(\mathbf{Y}_i \in t)$ with $P(\tilde{\mathbf{Y}}_i \in t)$. Therefore, by assuming that the feature variables are shifted by the function $\boldsymbol{\xi}(\mathbf{x};\Theta)$, we can apply our population-based formulation of the BFOS approach to the covariate adjusted feature data in order to construct a tree $T'^{\text{adj}(\mathbf{x})}$ that suitably adjusts for covariate effects.

Based on Result 1, the following facts regarding $T'^{\text{adj}(\mathbf{x})}$ hold:

1. Let $\tilde{Y}_\nu$ and $Y_{\nu,\mathbf{x}}$ correspond to any of the P features in $\tilde{\mathbf{Y}}$ and $\mathbf{Y}_\mathbf{x}$ (the group membership of which is unknown). We then have that the split $\tilde{Y}_\nu \le \tilde{c}_\nu$ in $T'^{\text{adj}(\mathbf{x})}$ is equivalent to the split $Y_{\nu,\mathbf{x}} \le \tilde{c}_\nu + \xi_\nu(\mathbf{x}|\boldsymbol{\theta}_\nu)$ in the tree based on $\mathbf{Y}_\mathbf{x}$. For example, if a feature that has been adjusted for some covariate effect is greater than a constant value for a specific split in $T'^{\text{adj}(\mathbf{x})}$, then this feature will be greater than a value that depends on that covariate when expressed on the original scale.
2. Regardless of the value of $\mathbf{x}$, $T'^{\text{adj}(\mathbf{x})}$ allows for the identification of a covariate independent subset of features that best characterizes the differences among the g groups while accounting for all relevant covariate effects.

4.3. *COVACT estimation using training data*

The first step in estimating the CDFs $F_{\tilde{\mathbf{Y}}|i}(.)$ needed to construct our covariate adjusted tree is to estimate the parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ from the available training data $(\mathbf{y}_{ij}, \mathbf{x}_{ij})$ $(i = 1, \ldots, g; j = 1, \ldots, n_i)$. Based on Tu *and others* (1997), a simple approach to do so is to assume that the conditional mean of the random feature vector $\mathbf{Y}_{ij} = (Y_{ij,1}, \ldots, Y_{ij,P})'$ is given by $E[\mathbf{Y}_{ij}|\mathbf{x}_{ij}] = \boldsymbol{\lambda}_i + \boldsymbol{\xi}(\mathbf{x}_{ij}; \Theta) = \boldsymbol{\lambda}_i + \left(\xi_1(\mathbf{x}_{ij}|\boldsymbol{\theta}_1), \ldots, \xi_P(\mathbf{x}_{ij}|\boldsymbol{\theta}_P)\right)'$, where $\boldsymbol{\lambda}_i = (\lambda_{1,i}, \ldots, \lambda_{P,i})'$ denotes the effect of belonging to the i^{th} group. We then use least squares (LS) estimation to obtain the estimates $\hat{\lambda}_{1,i}, \ldots, \hat{\lambda}_{P,i}, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ that minimize the criteria $Q_p = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij,p} - \lambda_{p,i} - \xi_p(\mathbf{x}_{ij}|\boldsymbol{\theta}_p))^2$ $(p = 1, \ldots, P)$. We note that there may be instances where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ are not identifiable, in which case $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$ are not unique. However, regardless of these estimates, the trees obtained from our construction approach produce the same results (details provided in supplementary material available at *Biostatistics* online). As noted below, the estimates of $\lambda_{1,i}, \ldots, \lambda_{P,i}$ are not used in our construction approach, so whether or not these parameters are identifiable does not affect our resulting tree.

Once we obtain the estimates $\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_P$, we can compute the covariate adjusted feature data $\hat{\tilde{\mathbf{y}}}_{ij} = (\hat{\tilde{y}}_{ij,1}, \ldots, \hat{\tilde{y}}_{ij,P})' = (y_{ij,1} - \xi_1(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_1), \ldots, y_{ij,P} - \xi_P(\mathbf{x}_{ij}|\hat{\boldsymbol{\theta}}_P))' = \mathbf{y}_{ij} - \boldsymbol{\xi}(\mathbf{x}_{ij}; \hat{\Theta})$. 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 $F_{\tilde{\mathbf{Y}}|i}(.)$, so that we can compute the empirical CDFs $\hat{F}_{\tilde{\mathbf{Y}}|i}(\tilde{\mathbf{c}}) = \sum_{j=1}^{n_i} I(\hat{\tilde{y}}_{ij,1} \le \tilde{c}_1, \ldots, \hat{\tilde{y}}_{ij,P} \le \tilde{c}_P)/n_i$. Based on the estimates $\hat{P}(\tilde{\mathbf{Y}}_i \in t)$ obtained from $\hat{F}_{\tilde{\mathbf{Y}}|i}(.)$, we can construct our covariate adjusted tree by applying the standard non-parametric BFOS approach to the covariate adjusted feature data. Thus, it is important to observe that our COVACT construction approach can be easily implemented using standard classification tree software and its variations.

In this manner, we can view our tree construction approach as a semi-parametric generalization of the approach developed by L&T. Specifically, we first use a parametric model to estimate the parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_P$ that we use to compute our covariate adjusted feature data. The CDFs for these adjusted feature data are then estimated non-parametrically in order to yield our final covariate adjusted tree.

5. APPLICATION TO POSTMORTEM TISSUE DATA

Our COVACT methodology is now illustrated in the context of four post-mortem brain tissue studies conducted by Sweet *and others* (2003, 2004, 2007, 2008) at the CCNMD. Across these studies, six biomarkers are examined, consisting of three measures taken from the primary auditory cortex: synaptophysin-immunoreactive (SY-IR) puncta density for Brodmann's areas (BA) 41 and for BA 42; pyramidal cell somal volume for BA 41 and for BA 42 (natural logarithm scale), and spinophilin-immunoreactive (SP-IR) puncta density for BA 41 and for BA 42. In each study, age at death, gender, PMI, and brain tissue storage time are included as covariates for both schizophrenia subjects and normal controls. Although there are slight differences in the numbers of subjects used in each study, there are 30 subjects common to all four studies that we use in our application. Our objective here is to illustrate how COVACT

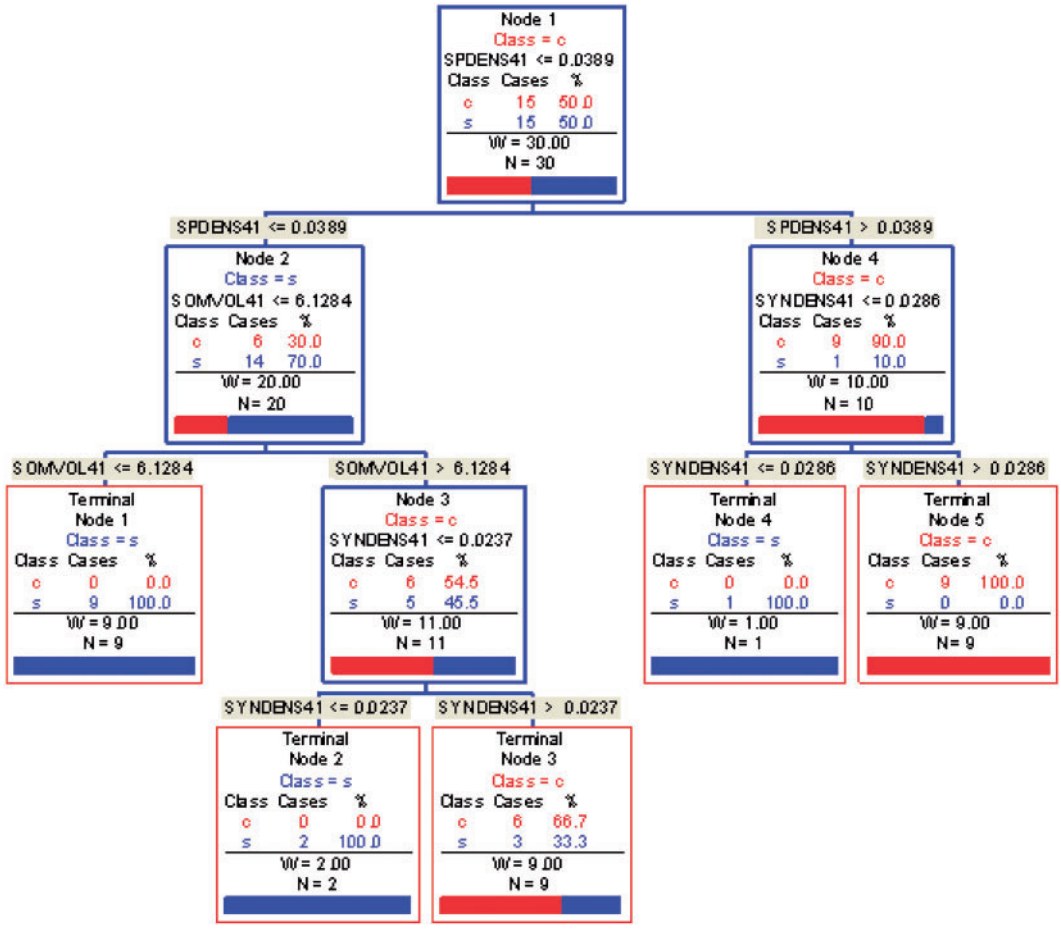

Fig. 1. COVACT for sweet data. SpDens41, SomVol41, and SynDens41 correspond to SP-IR puncta density, somal volume, and SY-IR puncta density for BA 41, adjusted for the effects of age at death, gender, PMI, and storage time. In each terminal node, a label of "class=s" corresponds to subjects classified as having schizophrenia, while that of "class=c" corresponds to subjects classified as controls.

and the traditional BFOS approach perform in identifying which subsets of biomarkers best characterize the differences between schizophrenia and control subjects. Performance is measured by estimating the misclassification rates from the resulting trees. The manner in which we applied COVACT, as well as the traditional BFOS approach and covariate adjusted LDA (for comparative purposes), is detailed in supplementary material available at *Biostatistics* online.

Figures 1 and 2 display, respectively, the trees produced from the biomarker data adjusted for the effects of the included covariates using COVACT, and the unadjusted biomarker data. In each terminal node of each tree, a label of "class=s" or "class=c" means that subjects with measurements falling into that node are classified into the schizophrenia or control groups, respectively. Adjusting for the effects of age at death, gender, PMI, and tissue storage time, COVACT identifies SY-IR puncta density, somal volume, and SP-IR puncta density for BA 41 as best characterizing the differences between schizophrenia subjects and normal controls, where the resulting tree is observed to misclassify only 10%

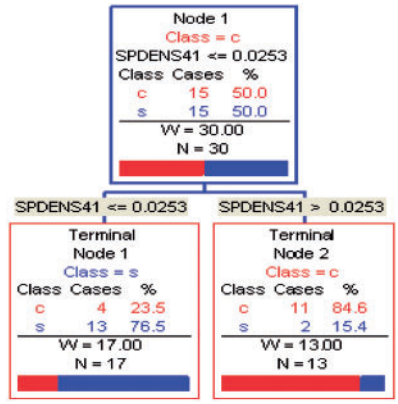

Fig. 2. Traditional classification tree for sweet data. In each terminal node, a label of "class=s" corresponds to subjects classified as having schizophrenia, while that of "class=c" corresponds to subjects classified as controls. SpDens41 corresponds to SP-IR puncta density for BA 41.

Table 1. *Classification results for sweet data* (*C—control, S—schizophrenia*)

Approach	True Group	Classified as C	Classified as S	Total	Observed error rate
COVACT	C	15	0	15	**0.1**
	S	3	12	15	
	Total	18	12	30	
Covariate adjusted LDA	C	10	5	15	**0.233**
	S	2	13	15	
	Total	12	18	30	
Unadjusted	C	11	4	15	**0.2**
	S	2	13	15	
	Total	13	17	30	

of all subjects (see Table 1). On the other hand, when these covariate effects are ignored, the traditional BFOS approach only identifies SP-IR puncta density for BA 41 and is observed to misclassify 20% of all subjects. The increased classification accuracy under COVACT, compared with the traditional BFOS approach, makes sense intuitively. COVACT removes the masking effects of covariates from the examined features, which would help to more accurately identify subsets of features that best characterize the group differences. In addition, when we apply LDA to the six biomarkers adjusted for the examined covariate effects, the resulting discriminant function misclassifies 23.3% of all subjects. Therefore, COVACT also outperforms covariate adjusted LDA with respect to classification accuracy. Compared with LDA, classification trees allow for a finer and more flexible partitioning of the feature space. Therefore, it is easier for classification trees to uncover the structure of the feature data that truly defines the differences among groups.

Due to the small sample size, however, this difference in classification accuracy needs to be viewed with caution.

6. Simulation study

6.1. *Motivation*

To more extensively compare the classification performances of COVACT and the traditional BFOS approach, we conduct a simulation study based on data settings that mimic those examined by the four Sweet *and others* studies discussed in Section 5. Specifically, we generate feature values using the six biomarkers measured across these studies and covariate values using the values for subject's age at death, PMI, and brain tissue storage time (averaged across all four studies). In conducting our study, we are interested in evaluating how the degree of correlation between features and covariates impacts classification accuracy for COVACT relative to traditional classification trees and, thus, only consider the continuous covariates of subject's age at death, PMI, and brain tissue storage time. We compare the classification accuracy of both approaches in seven scenarios where in each, we consider the cases of $P = 2, 3, 4,$ and 6 features with $J = 30, 60,$ and 240 observations are considered and the generated data satisfy the main assumption of COVACT described in Section 4.2. The manner in which this study was conducted is detailed in supplementary material available at *Biostatistics* online.

6.2. *Conclusions*

The classification results from our simulation study are detailed in supplementary material available at *Biostatistics* online. In short, both COVACT and the traditional BFOS approach have lower misclassification rates, on average, as the number of features increases. However, COVACT yields generally lower misclassification rates relative to the traditional BFOS approach, regardless of the number of features, observations, and covariates that are adjusted for. This also holds in each group we consider, which indicates that COVACT is generally less conservative than the traditional BFOS approach. In addition, the decrease in misclassification rates for COVACT relative to the traditional BFOS approach becomes more pronounced as the number of covariates increases. Therefore, these results, in conjunction with those from Section 5, these results provide support for the improved classification performance of COVACT.

7. Discussion

We develop our proposed COVACT methodology by first approaching the traditional BFOS tree construction method, which is primarily meant to be used with training data, from a population-based standpoint. In doing so, we not only provide a parametric alternative to the standard non-parametric BFOS approach but we also establish a basis to integrate the conditional distribution of the feature data for a fixed covariate value into the BFOS approach, in order to account for the relationship that may exist between the features and covariates of interest. Considering that our proposed modification for the BFOS method may produce trees that vary depending on covariate value, we formulate our COVACT construction approach, which has two desirable properties. First, and perhaps most importantly, it allows us to use the traditional BFOS approach to construct a tree using feature data that have been adjusted for covariate effects, so that available software packages can be used. Also, COVACT produces a tree that can be used to identify a unique subset of features, independent of covariate value, that best characterizes the differences among the groups of interest while adjusting for the effects of all relevant covariates. Through simulation studies and a practical application, we have shown that when covariate effects are evident, COVACT can yield more accurate results than traditional classification trees that ignore these effects.

With COVACT, we have developed a tree-based approach that accounts for covariate effects when identifying which subset of features best characterizes the differences among groups, a subset that does not change depending on covariate value. However, as was pointed out in Sections 1 and 3.3, there may be situations where it is meaningful to have this subset of features vary depending on a particular covariate

value(s), for example, subject's age at death in the context of post-mortem tissue studies. There may also be situations in which covariate effects on the examined features are differentially expressed across groups, a case in which the main assumption of the linear invariance property on which COVACT is based is violated. In these instances, one will need to consider approaches for constructing covariate-dependent conditional classification trees with available training data, which are discussed in Asafu-Adjei (2011).

We are familiar with studies in the literature, for example, Breiman (2001b), that discuss classification methods with higher classification accuracy than classification trees, including the ensemble learning approach of Random Forests. However, these same studies demonstrate the great decrease in simplicity and interpretability for these classification methods relative to classification trees. Although we place a high premium on accuracy, we also place a high premium on the interpretability of our results. Therefore, we believe that the high level of interpretability for classification trees goes a long way in compensating for the relatively small decrease in accuracy, which is why we develop our adjustment approach in the context of classification trees.

One variation to our setting that is of interest is unequal misclassification costs, which we can easily incorporate into our COVACT methodology. In addition, there may be studies where subjects are not only measured on covariates, but are also matched across different groups on several demographic factors. Along with covariates, group matching may also have a significant impact on the features of interest, in which case a subtle modification to our COVACT approach is required (see Asafu-Adjei, 2011, for a detailed discussion).

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

We thank Dr. Robert A. Sweet (Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA) for providing the postmortem brain tissue data used in our application. We gratefully acknowledge the efforts of Dr. David A. Lewis (Departments of Neuroscience and Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA), and the research staff of the Conte Center for the Neuroscience of Mental Disorders. The authors also thank the reviewers and associate editor of *Biostatistics* for their helpful suggestions. *Conflict of Interest:* None declared.

FUNDING

This work was supported by the National Institutes of Health (T32NS048005, F32NS081904, and MH084053) which supported J.K.A.; (MH084053 and MH103204) which supported A.R.S. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

REFERENCES

ASAFU-ADJEI, J. (2011). Covariate adjusted discrimination with applications to neuroscience, [PhD. Thesis]. Pittsburgh, PA, USA: University of Pittsburgh.

ASAFU-ADJEI, J. K., SAMPSON, A. R., SWEET, R. A. AND LEWIS, D. A. (2013). Adjusting for matching and covariates in linear discriminant analysis. *Biostatistics* **14**, 779–791.

BREIMAN, L. (2001a). Random forests. *Machine Learning* **45**, 5–32.

BREIMAN, L. (2001b). Statistical modeling: The two cultures. *Statistical Science* **16**, 199–231.

BREIMAN, L, FRIEDMAN, J. H., OLSHEN, R. A. AND STONE, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth Int. Group.

COCHRAN, W. G. AND BLISS, C. I. (1948). Discriminant functions with covariance. *Annals of Mathematical Statistics* **19**, 151–176.

CORTES, C AND VAPNIK, V. (1995). Support-vector networks. *Machine Learning* **20**, 273–297.

CRAINICEANU, C. M., DOMINICI, F AND PARMIGIANI, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* **95**, 635–651.

DUKART, J, SCHROETER, M. L., MUELLER, K, Alzheimers Disease Neuroimaging Initiative. (2011). Age correction in dementia–matching to a healthy brain. *PLoS ONE* **6**, e22193.

FRISTON, K. J., HOLMES, A. P., WORSLEY, K. J., POLINE, J. P., FRITH, C. D. AND FRACKOWIAK, R. S. J. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping* **2**, 189–210.

KIM, H AND LOH, W. Y. (2003). Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics* **12**, 512–530.

KNABLE, M. B., BARCI, B. M., BARTKO, J. J., WEBSTER, M. J. AND TORREY, E. F. (2002). Molecular abnormalities in the major psychiatric illnesses: Classification and regression (CRT) analysis of post-mortem prefrontal markers. *Molecular Psychiatry* **7**, 392–404.

LACHENBRUCH, P. A. (1977). Covariance adjusted discriminant functions. *Annals of the Institute of Statistical Mathematics* **29**, 247–257.

LI, Y., GONG, S. AND LIDDELL, H. (2003). Recognising trajectories of facial identities using kernel discriminant analysis. *Image and Vision Computing* **21**, 1077–1086.

NONYANE, B. A. S. AND FOULKES, A. S. (2008). Application of two machine learning algorithms to genetic association studies in the presence of covariates. *BMC Genetics* **9**, 71.

SHANG, N. AND BREIMAN, L. (1996). Distribution based trees are more accurate. In: *Proceedings of the International Conferance on Neural Information Processing, Hong Kong*, **1**, 133–138.

SWEET, R. A., BERGEN, S. E., SUN, Z., MARCSISIN, M. J., SAMPSON, A. R. AND LEWIS, D. A. (2007). Anatomical evidence of impaired feedforward auditory processing in schizophrenia. *Biological Psychiatry* **61**, 854–864.

SWEET, R. A., BERGEN, S. E., SUN, Z., SAMPSON, A. R., PIERRI, J. N. AND LEWIS, D. A. (2004). Pyramidal cell size reduction in schizophrenia: Evidence for involvement of auditory feedforward circuits. *Biological Psychiatry* **55**, 1128–1137.

SWEET, R. A., HENTELEFF, R. A., ZHANG, W., SAMPSON, A. R. AND LEWIS, D. A. (2008). Reduced dendritic spine density in auditory cortex of subjects with schizophrenia. *Neuropsychopharmacology* **34**, 374–389.

SWEET, R. A., PIERRI, J. N., AUH, S., SAMPSON, A. R. AND LEWIS, D. A. (2003). Reduced pyramidal cell somal volume in auditory association cortex of subjects with schizophrenia. *Neuropsychopharmacology* **28**, 599–609.

TU, X. M., KOWALSKI, J., RANDALL, J., MENDOZA-BLANCO, J., SHEAR, M. K., MONK, T. H., FRANK, E. AND KUPFER, D. J. (1997). Generalized covariance-adjusted discriminants: Perspective and application. *Biometrics* **53**, 900–909.

WANG, C., PARMIGIANI, G. AND DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68**, 661–686.

WILSON, A. AND REICH, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics* **70**, 852–861.

[*Received October 21, 2015; revised February 24, 2017; accepted for publication March 16, 2017*]