Statistics and Its Interface Volume 1 (2008) 169–178

# A tree-based method for modeling a multivariate ordinal response[*]

Heping Zhang and Yuanqing Ye

Motivated by a real example of understanding the so-called "building related occupant complaint syndromes" (BROCS), we propose a tree-based method for analyzing a multivariate ordinal response. Our method is semiparametric by assuming a within-node parametric distribution on the adaptive nonparametric tree framework. We use simulation experiments to demonstrate the ability of our method to identify underlying structures in the data and the fact that analyzing ordinal response data with proper methods that take ordinality into account is considerably more powerful than dichotomization. The reanalysis of the BROCS data also suggests new insights that go beyond a previous analysis based on the dichotomization.

AMS 2000 subject classifications: Primary 62G08; secondary 62P12.

Keywords and phrases: Classification trees, Ordinal variables, Multivariate model, Working environment, Respirotary symptoms.

## 1. INTRODUCTION

In many studies such as health surveys, researchers collect information from multiple items or sources to determine the condition or disease status of a subject. For example, in a study of the so-called "building related occupant sick syndromes" (BROCS), responses on six clusters of complaints were collected from each of the 6,800 respondents. It is noteworthy that the choices for many items in the BROCS study are "never", "rarely", "sometimes", "often", and "always", or a subset of these choices. To analyze such data, a common practice is to dichotomize the ordinal scaled responses by collapsing them into "yes" or "no" and then model the derived multivariate binary response as in Zhang (1998). While convenient, such a dichotomization is not ideal because it leads to loss of information and the decision of collapsing which levels into the same group is often subjective. We will demonstrate later that analyzing ordinal response data with proper methods that take ordinality into account is considerably more powerful than dichotomization.

In the past decade, it has become a very active research area to directly analyze ordinal responses instead of dichotomizing them first. Agresti (1999) provides a thorough review and extensive literature on some of the advances in this area. For example, Hedeker and Gibbons (1994) proposed a random-effect regression model for the analysis of multiple correlated ordinal responses. They assumed a linear regression model for latent unobserved continuous responses that underlie the observed ordinal responses. Then, either a cumulative probit or cumulative logit link function was used to model the response functions. Ten Have (1996) proposed a mixed effects model to analyze ordinal responses and derived a closed form for the marginal log-likelihood when the distribution of the random effects is log-gamma with the cumulative complementary log-log link function.

It is evident from Agresti (1999) and other work that random or mixed effects models are the basis of the parametric models for analyzing ordinal responses and that many useful models and approaches have been established. However, little progress has been made in using nonparametric techniques to deal with correlated ordinal responses, which can be particularly useful when the underlying structure in the data is not apparent and an exploratory step is desirable. In one exception, Kauermann (2000) introduced local smoothing in a regression model for clustered or longitudinal data with ordinal responses. In essence, the nonparametric technique is used to estimate the marginal parameters such as fixed effects and within-subject correlations. The main motivation for this work is to fill in this methodological gap.

Analyzing ordinal response data with proper methods that take ordinality into account is considerably more powerful than dichotomization (e.g., Zhang et al. 2003, Zhang et al. 2006). We mentioned a few different ways to analyze a multivariate ordinal response. In this paper, we develop a tree-based approach to the analysis of multivariate ordinal data, because it is an undeveloped area, and in many studies we have little information as to which and how predictors are related to the responses.

Like Zhang (1998) and Kauermann (2000), our approach is a hybrid of nonparametric and parametric methods, and is best termed as a semiparametric one. The idea is that the specific tree structure is not known and will be determined from the data – this is the nonparametric aspect of the method – and that whenever we have a tree, even temporarily, we assume a particular form of within-node distribution – this is the parametric aspect of the method.

[*]This research was supported in part by NIDA grants K02DA017713 and R01DA016750.

This idea has been frequently used in the development of tree-based methods (Zhang, 1998; Zhang and Singer 1999). Later, we will describe in detail what we mean by a node as well as other terminologies.

# 2. METHOD

## 2.1 The model

Assume that there are $N$ units of observations in the data. Within unit $i$, we observe $n$ ordinal responses denoted by $z_i$ and $p$ covariates denoted by $x_i$. In longitudinal studies, the set of responses within a unit results from repeated measures from the same subject, and the covariates may be constant in the unit or depend on the component of the unit, i.e., time varying. In other studies such as in Zhang (1998), the set of responses within a unit is comprised of different sites or symptoms of a disease, e.g., different sites of cancer, different respiratory symptoms, comorbid psychiatric diseases, and different substance abuses. Again, the covariates can be constant within the unit, or vary by the individual component of the unit. In this work, we only consider covariates that are common within the observation unit.

Let $z_{ij}$ denote the $j$th ordinal response in the $i$th unit, and suppose that it takes a value from $1, \ldots, K$. For clarity and without loss of generality, the number of categories $K$ is assumed to be the same for all response variables, because we can create extra levels with zero frequency when different responses have different numbers of categories. Let us define $K-1$ indicator variables $y_{ijk} = I(z_{ij} > k)$, for $k = 1, \ldots, K-1$. Here $I(\cdot)$ is the indicator function. Let

$$(1) \qquad \begin{aligned} y_{ij} &= (y_{ij,1}, \ldots, y_{ij,K-1})^{\mathrm{T}}, \\ y_i &= (y_{i1}^{\mathrm{T}}, \ldots, y_{in}^{\mathrm{T}})^{\mathrm{T}}, \end{aligned}$$

Then, the observed responses from the $i$th unit can be rewritten as

$$y_i = (y_{i1,1}, \ldots, y_{i1,K-1}, \ldots, y_{in,1}, \ldots, y_{in,K-1})^{\mathrm{T}},$$

and the covariates are

$$x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}},$$

$i = 1, \ldots, N$. Our objective is to characterize the distribution of $y_i$ based on $x_i$.

## 2.2 Node-split criterion

As in the general tree methodology (Breiman et al. 1984 and Zhang and Singer 1999), the construction of a tree begins with splitting the root node that consists of all data in a learning sample. Once the root node is divided into two daughter nodes, we apply the same procedure to partition the daughter nodes. This process is referred to as the recursive partitioning procedure, and it is the thrust of the tree formation. Thus, in this section, we concentrate on splitting the root node into the left daughter($t_L$) and right daughter($t_R$).

For an ordered covariate (on an ordinal or continuous scale), say $x_j$, and a threshold value $c$, whether unit $i$ is assigned to the left or right daughter node is based on the question whether or not $x_{ij} > c$. If $x_j$ is a categorical covariate, the assignment is made according to a particular set of some levels that this covariate can take. All covariates and their splitting values together usually yield many ways of splitting the root node. Hence, a splitting criterion to select the best split is needed.

Following the approach of Zhang (1998) for handling multiple binary responses, we assume an intermediate within-node joint distribution of the responses within a unit. Precisely, for the $i$th unit, we consider a within-node joint distribution for $z_i$ as follows

$$(2) \qquad f_i(z_i; \psi, \omega) = \exp\{\psi^{\mathrm{T}} y_i + \omega^{\mathrm{T}} w_i - A_i(\psi, \theta)\},$$

where $\psi$ and $\omega$ are vectors of parameters, $A_i(\psi, \omega)$ is the normalization constant which is the summation over $K^n$ possible assignments for the value of $z_i$. Moreover, $y_i$ is defined in (1) and

$$(3) \qquad \begin{aligned} w_i = (&y_{i1,1}y_{i2,1}, \ldots, y_{i1,1}y_{i2,(K-1)}, \ldots, \\ &y_{i(n-1),(K-1)}y_{in,(K-1)})^{\mathrm{T}}. \end{aligned}$$

Because $w_i$ consists of pairwise interactions between all indicators of two responses, its coefficient vector $\omega$ contains the information on the "correlation" of the two responses. In fact, in simple cases, these coefficients are odds ratios, which are commonly used to describe the association between two binary variables.

Without restrictions, $\psi$ and $\omega$ have $(K-1)n$ and $(K-1)^2 n(n-1)/2$ elements, respectively. To reduce the number of the free parameters and retain the ordinality property of the responses, we introduce a link matrix $C$ and let $\omega = C\theta$, where

$$(4) \qquad C = \mathbf{1}_n \otimes I_{(K-1)^2}.$$

Here $I_m$ denotes an $m \times m$ identity matrix for an integer $m$, $\mathbf{1}_n$ is an $n$-vector of ones, and $\otimes$ refers to the operator for the Kronecker product. Then, the joint probability distribution of $z_i$ becomes

$$(5) \qquad f_i(z_i; \psi, \theta) = \exp\{\psi' y_i + (C\theta)' w_i - A_i(\psi, \theta)\}.$$

**Remark.** The implication of this parameter reduction is to assume that the correlation structure among any two pair of responses is the same, but varies for different ordinal levels to retain the ordinality of the responses. We should note that the heterogeneity in different responses is reflected through the "prevalence" parameter $\psi$. In other words, the "prevalence" parameters depend on a specific response and its ordinal level while the "correlation" parameters are only functions of the ordinal level. Such assumptions may not be ideal, but are quite common in analysis of correlated data.

Once the within-node distribution is defined, the maximum log-likelihood can serve as a splitting criterion. That is, for any node $t$, we obtain

$$h(t) = \sum_{\{\text{unit } i \in t\}} \{\hat{\psi}' y_i + (C\hat{\theta})' w_i - A_i(\hat{\psi}, \hat{\theta})\}, \tag{6}$$

where $\hat{\psi}$ and $\hat{\theta}$ are the maximum likelihood estimators of $\psi$ and $\theta$ based on samples in node $t$. Thus, we select a split from all allowable ones to maximize the sum of the generalized entropies, $h(t_L)+h(t_R)$, from the resulting two daughter nodes. Because the pool of allowable splits is usually large and the assessment of each allowable split entails a maximization, we want to reduce the computational burden to its minimal level. In the appendix, we discuss in detail the computation issues.

In addition to the distribution assumed in (2), there are other alternative distributions that are commonly used in analysis of multivariate ordinal variables. For example, we considered a simple approach by treating the ordinal responses as numerical responses, and we will present simulation results from this approach below. In addition, we evaluated another approach based on the multivariate distribution of Hedeker and Gibbons (1994). This approach is computationally complicated and the resulting trees were not preferable according to our simulation (data are not presented here). On the basis of this empirical evidence, we adopted the exponential distribution in (2).

## 2.3 Tree pruning

So far, we have explained the recursive partitioning procedure for growing a tree. Because we do not employ any stopping procedure, the resulting tree is generally large and over grown. Thus, it is important to determine an appropriate size of the final tree for interpretation. This step is usually referred to as pruning. The commonly used procedures include error based pruning (Quinlan 1987, 1993), minimum error pruning (Niblett and Bratko 1986), critical value pruning (Mingers 1987), and cost-complexity pruning (Breiman 1984). Esposito et al. (1997) compared these pruning methods and found no clear winner. Later, Li et al. (2001) proposed a pruning method that combines dynamic programming and classical optimization technique. Here, we use the cost-complexity concept, because it can be easily generalized from a single response to multiple responses.

Briefly, we divide the entire sample into a number of groups (e.g., 2, 5 or 10) of approximately equal sizes. If we use a $v$-fold cross validation, the original sample is randomly divided into $v$ groups. We repeat the tree growing procedure $v$ times, and each time we use $v-1$ of the $v$ subgroups of the sample as the learning sample alternatively for tree growing and the left-over subgroup as the test sample to evaluate the predictive quality of the tree. The final tree is determined to optimize the predictive quality based on the cross-validation.

To determine the final tree, we follow the steps of Breiman et al. (1984). There are four preparation steps before cross validation is applied. First, we introduce a complexity parameter, $\alpha$ that is used to penalize a large tree. Secondly, let $\tilde{T}$ denote the set of all terminal nodes of a tree $T$. We define the cost for tree $T$, denoted by $R(T)$, as

$$R(T) = -\sum_{t \in \tilde{T}} \sum_{\text{unit } i \in t} \log f(y_i; \hat{\Psi}, \hat{\theta}), \tag{7}$$

where $f$ is defined as in (5). Thirdly, we introduce a cost-complexity for a tree $T$ as

$$R_\alpha(T) = R(T) + \alpha |\tilde{T}|, \tag{8}$$

where $|\tilde{T}|$ is the number of terminal nodes. Lastly, we can find an increasing sequence of complexity parameters,

$$\alpha_0 < \alpha_1 < \cdots < \alpha_m, \tag{9}$$

each of which leads to a smallest subtree of $T$ with the lowest cost. Here, $\alpha_0 = 0$. It is important to note that the resulting sequence of subtrees is nested. That is, we have a sequence of nested optimal subtrees as follows:

$$T_0 \supset T_1 \supset \cdots \supset T_m,$$

where $T_0$ is the initial tree and corresponds to $\alpha_0$ and $T_n$ is the smallest tree containing the root node only.

Now, we are ready to select the final tree from the above sequence of optimal subtrees through cross validation. During the $k$th fold of the cross validation, we regrow a large tree, denoted by $T_0^{(k)}$ using the learning sample. Then, we apply the sequence of the complexity-parameters in (9) to produce a nested sequence of subtrees of $T_0^{(k)}$. We denote this sequence by

$$T_0^{(k)} \supseteq T_1^{(k)} \supseteq \cdots \supseteq T_m^{(k)}.$$

For each subtree $T_l^{(k)}$, we compute its cost, $R^{cv}(T_l^{(k)})$, as in (7) using the test sample, whereas $\hat{\Psi}$ and $\hat{\theta}$ are fixed from the learning sample. The cross validation cost, $R^{cv}(T_l)$, of subtree $T_l$ is the average of $R^{cv}(T_l^{(k)})$ over the $v$ folds. We can select the subtree with the smallest cross validation cost. However, as discussed in Breiman et al. (1984), it is sometimes helpful to consider the so-called one-standard error rule to overcome the instability of the subtree with the minimal cross validation cost. Instead of choosing the subtree $T_s$ with the minimal cross validation cost, we choose the smallest subtree $T_{s_1}$ such that

$$R^{cv}(T_{s_1}) \leq R^{cv}(T_s) + SE(R^{cv}(T_s)), \tag{10}$$

where $SE(R^{cv}(T_s))$ is the standard error of $R^{cv}(T_s)$.

Theoretical estimation of $SE(R^{cv}(T_s))$ is complex even in the simplest tree framework. Thus, we repeat the five-fold cross-validation process a number of times (e.g., 10) and obtain an estimate of $SE(R^{cv}(T_l))$ for $l = 0, \ldots, m$.

## 3. SIMULATION

In this section, we present simulation studies to evaluate the node-splitting measure, $h(t)$, in (6). The data are simulated as follows. We included the same 22 covariates as in the real data set to be presented in Section 4. A latent continuous variable $U$ underlying the observed ordinal variable $Z$ was generated from the following model:

$$u_{ik} = b_i + 0.5I(x_{15} > 1) + 1.0I(x_{13} > 3) + 0.5I(x_7 > 1) + \epsilon_{ik}, \tag{11}$$

where $b_i \sim N(0, 1)$ is the random intercept, $\epsilon_{ik} \sim N(0, 2)$ is the residual term, and they are all independent for $k = 1, \ldots, 6$, $i = 1, \ldots, 6800$. A 3-level ordinal response is determined by

$$z_{ik} = \begin{cases} 0 & \text{if } u_{ik} \leq 1, \\ 1 & \text{if } u_{ik} \in (1, 3], \\ 2 & \text{if } u_{ik} > 3. \end{cases}$$

Thus, 19 of the 22 covariates were included in the dataset as noise variables. We should note that the data generating model is not in a tree structure.

To speed up the computation, we used a 2-fold cross-validation during the tree pruning, and replicated the cross-validation 10 times to obtain the standard error of the cross validation estimate of the tree cost. The simulation experiment was replicated 100 times.

The resulting trees correctly selected $X_7 > 1$, $X_{13} > 3$, and $X_{15} > 1$ as the splits in 96 of the 100 runs. $X_{13} > 3$ is always the one to split the root node. In only 1 of the 100 runs, the tree included an additional split based on $X_{10} > 1$. In only 3 of the 100 runs, the trees did not select $X_{15} > 1$. Thus, the trees based on $h(t)$ are highly sensitive and specific in revealing the true splits.

As mentioned earlier, we also evaluated alternative splitting criteria. For example, the simplest way is to treat the ordinal responses as numerical values and use the correlation matrix with each node as the basis for splitting. This criterion was studied by Segal (1992) and Zhang (1998). In 21 of the 100 runs, the final trees selected a noise variable. In six runs out of 100, the final trees had only one split, which was $X_{13} > 3$. The other two variables were not used. Hence, treating the ordinal responses as if they were numerical is not a satisfactory approach.

To further evaluate the validity of our method, we performed additional simulations to investigate the ability of our proposed method in identifying the relevant splits when the data are generated from tree-based models. Although the real data are not likely to conform the exact tree structures, it is still useful to understand the performance of the trees under the hypothetical situations.

Our simulation data were generated from a tree with 3 layers and 4 terminal nodes as displayed in Figure 1. The units in the four terminal nodes had different joint probability distributions determined by the parameters $(\psi_{11}, \psi_{12}, \psi_{21}, \ldots, \psi_{62}, \theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$. These parameters were set to $(-1., -3., -1.4, -3., -1., -3.2, -1.2, -3., -1.4, -3.6, -1.4, -3.2, 0.5, 0.0, 0.4, 0.5)$; $(-1., -3., -1.5, -3., -1., -3.2, -1.2, -3., -1.2, -3.2, -1.3, -3.1, 0.5, 0.4, 0.4, 0.5)$; $(-1., -3.1, -0.9, -3., -1.1, -3.1, -1.0, -3.2, -1.0, -3.2, -1.0, -3.4, 0.5, 0.2, 0.4, 0.5)$; and $(-1., -3.1, -1.0, -3., -1.4, -3.1, -1.0, -3., -1.1, -3.2, -1.0, -3.4, 0.5, 0., 0., 0.5)$ for nodes 4, 5, 6 and 7, respectively. Again, we used the same covariates as the real data set to be analyzed below, and 19 of the 22 covariates were included in the data sets as noise variables. Furthermore, we generated six 3-level ordinal responses.

We replicated our simulation 50 times. After pruning, 49 of the 50 final trees were exactly the same as in Figure 1. The lone missed one contained the three underlying splits, however. Thus, this simulation experiment validates our method.

If the goal is to identify relevant splits that are predictive of the responses, it is useful to know whether we can also achieve this goal by collapsing the three-level ordinal responses to two-level binary responses and applying the existing tree based method. For example, let us collapse the first two levels into the same one. Using the same 50 replications as described above, we constructed 50 trees using the multiple binary outcomes. Thirteen out of the 50 trees are the same as the underlying tree in Figure 1. Thirty trees out of 50 are of the same form as depicted in Figure 2. The remaining 7 trees are extensions of the tree shown in Figures 1 and 2. Therefore, certain information is lost as a result of the dichotomization while some essential information is also kept.

Using this experiment, we re-evaluated the idea of treating the ordinal responses as continuous ones. Table 3 summarizes all of the final tree structures. Like the dichotomization approach, this approach does not perform as well as the use of $h(t)$ in (6).

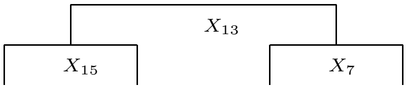


*Figure 1. Predefined Tree Used in the Simulation.*

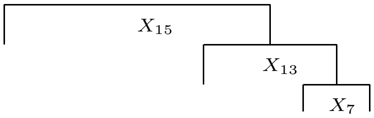


*Figure 2. Final Tree Identified by Most Simulation Runs Using Collapsed Binary Responses.*

## 4. APPLICATION

We now apply our method to the BROCS data set analyzed by Zhang (1998). BROCS contains a set of highly correlated symptoms of discomfort reported by occupants of buildings. It is a common occurrence in office buildings, hospitals, etc. all over the world. There are several common symptoms of BROCS such as irritation of the eyes, nose, and throat, headache, and nausea. Continuing efforts have been dedicated to understand the nature and find the cause of BROCS. See Zhang (1998) for the related references. Our data set is a subset of the data from a 1989 survey of 6,800 employees of the Library of Congress (LOC) and the headquarters of the Environmental Protection Agency (EPA) in the United States. In Zhang (1998), BROCS is represented by six binary outcome variables in Table 1, and 22 predictors described in Table 2 are considered as risk factors of BROCS. While we use the same covariates, because the six

*Table 1. Ordinal Response Variables*

| Response | Cluster | Symptoms |
|---|---|---|
| $y_1$ | CNS | Difficulty remembering/ concentrating, dizziness, lightheadedness |
| $y_2$ | Upper airway | Runny/stuffy nose, sneezing, cough, sore throat |
| $y_3$ | Pain | Aching muscles/joints, pain in back/shoulders/neck, pain in hands/wrists |
| $y_4$ | Flu-like | Nausea, chills, fever |
| $y_5$ | Eyes | Dry, itching, or tearing eyes; sore/strained eyes; blurry vision |
| $y_6$ | Lower airway | Wheezing in chest, shortness of breath, chest tightness |

*Table 3. Distinct Final Trees and Their Frequencies In 50 Simulation Runs by Treating Ordinal Responses as Continuous Measures*

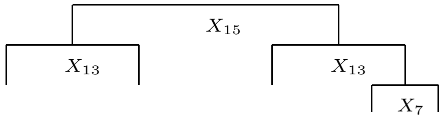
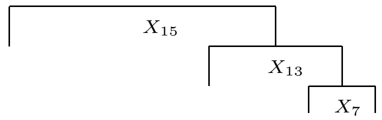
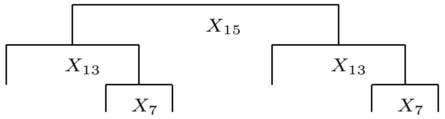
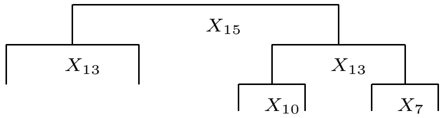
| Frequency | Tree Structure |
|---|---|
| 27 |  |
| 12 |  |
| 10 |  |
| 1 |  |

*Table 2. Description of Explanatory Variables*

| Name | Questions | Answer |
|---|---|---|
| $x_1$ | What is the type of your working space? | enclosed office with door, cubicle without door, stacks, etc. |
| $x_2$ | How is your working space shared | single, occupant, shared, etc. |
| $x_3$ | Do you have a metal desk? | yes or no |
| $x_4$ | Do you have new equipment at your work area? | yes or no |
| $x_5$ | Are you allergic to pollen? | yes or no |
| $x_6$ | Are you allergic to dust? | yes or no |
| $x_7$ | Are you allergic to mold? | yes or no |
| $x_8$ | How old are you? | 16–79 years old |
| $x_9$ | Gender | male or female |
| $x_{10}$ | Is there too much air movement at your work area? | never, rarely, sometimes, often, always |
| $x_{11}$ | Is there too little air movement at your work area? | same as $x_{10}$ |
| $x_{12}$ | Is your work area too dry? | same as $x_{10}$ |
| $x_{13}$ | Is the air too stuffy at your work area? | same as $x_{10}$ |
| $x_{14}$ | Is your work area too noisy? | same as $x_{10}$ |
| $x_{15}$ | Is your work area too dusty? | same as $x_{10}$ |
| $x_{16}$ | Do you experience glare at your workstation? | no, sometimes, often, always |
| $x_{17}$ | How comfortable is your chair? | reasonably, somewhat, very uncomfortable, no one specific chair |
| $x_{18}$ | Is your chair easily adjustable? | yes, no, not adjustable |
| $x_{19}$ | Do you have influence over arranging the furniture? | very little, little, moderate, much, very much |
| $x_{20}$ | Do you have children at home? | yes or no |
| $x_{21}$ | Do you have major childcare duties? | yes or no |
| $x_{22}$ | what type of job do you have? | managerial, professional, technical, etc. |

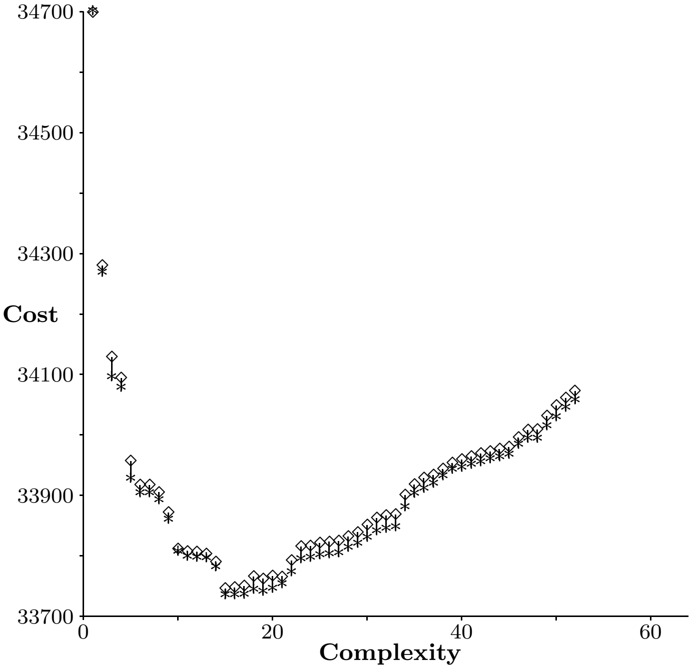


*Figure 3. Cost-Complexity for Sequence of Nested Subtrees. ***, Cross-Validation Estimates of Cost; ⋄ ⋄ ⋄, One Standard Error Above the Cross-Validation Estimated Cost.*

outcome variables were originally defined in ordinal scales, we reanalyze the data set using the six three-level ordinal responses (0: rarely; 1: sometimes; 2: often).

Based on (5), an initial tree $T_0$ of 103 nodes was grown. The pruning step produced a sequence of 52 nested optimal subtrees. The cost-complexity plot for these nested subtrees is displayed in Figure 3. Ten replications of 5-fold cross-validation were used to obtain the estimates of the subtree costs and their standard errors. As can be seen in Figure 3, the 1-standard error rule leads to the final subtree with 15 terminal nodes, which is displayed in Figure 4. This figure also summarizes the information for the 15 terminal nodes.

Figure 4 reveals more information than that presented by Zhang (1998). First, it includes all risk factors related to air quality measures that were identified by Zhang (1998). Furthermore, it suggests that subjects in terminal nodes 16 and 17 reported the fewest numbers of complaints for all six types of symptoms. People in those groups experienced better air quality in their working environment. For example, the air was never too dusty, rarely too dry, and might be occasionally too stuffy. Comparison of nodes 16 to 17, and 22 to 23 reveals the gender difference in that females complained more frequently of discomfort than males, especially for pain.

Even though subjects in terminal nodes 24 and 25 were allergic to mold, comparison of these nodes suggests that a too dusty air quality in the work space resulted in more frequent discomfort in the upper airway and pain.

It is interesting to note that Zhang (1998) did not note a clear increase of discomfort of eye symptoms as a result of experiencing a glare. Examining terminal nodes 26 and 27 suggests a notable increase of eye symptoms at level 2, although this is less notable at level 1. An increase of comfort in some of the other symptoms is more apparent.

Figure 4 reveals that the variation in the level 1 responses among all terminal nodes is relatively modest for every symptom. In other words, the 22 covariates are not so predictive at this discomfort level. The variation in the level 2 responses is evident, although as noted by Zhang (1998) the eyes-related complaints are quite rare in general. In contrast to Zhang (1998) in which levels 1 and 2 were merged, our analysis suggests that it would have been more reasonable to merge levels 0 and 1. In other words, if collapsing some levels are desirable, our proposed method can provide useful information as to how the collapsing should be executed.

## 5. DISCUSSION

We proposed a tree-based method to analyze multivariate ordinal responses. Our method is semiparametric and takes advantage of the adaptive and intuitive tree structures as well as the simplicity of parametric assumptions. While this general idea appears straightforward, developing a computationally feasible and practically interpretable approach remains a challenge.

It is very common for multivariate ordinal responses to have missing data. For example, in longitudinal analysis, missing values occur when the subject of interest is unavailable for the measurements. In our analysis, we assumed that the number of observations are equal for each subject. Since missing data in the response variables result in a varying number of observations for each subject, it remains important to investigate multiple imputation schemes in the context of tree based modeling.

While we have examined multivariate responses with the same number of ordinal scales, our model and method are readily applicable to a mix of ordinal and nominal responses, including the binary case. The challenge is to define a tractable and justifiable joint distribution.

Unlike parametric methods that are best suited for data with simple relations in the data, tree-based approach is more advantageous for data that include a large number of covariates and response variables, such as the BROCS data set. We have demonstrated through simulation studies that our method can perform exploratory data analysis, choose the important covariates from a long list of potentially influential covariates, and unravel the appropriate data structures. Applying our method to the BROCS data set led to useful insights into the understanding of the BROCS.

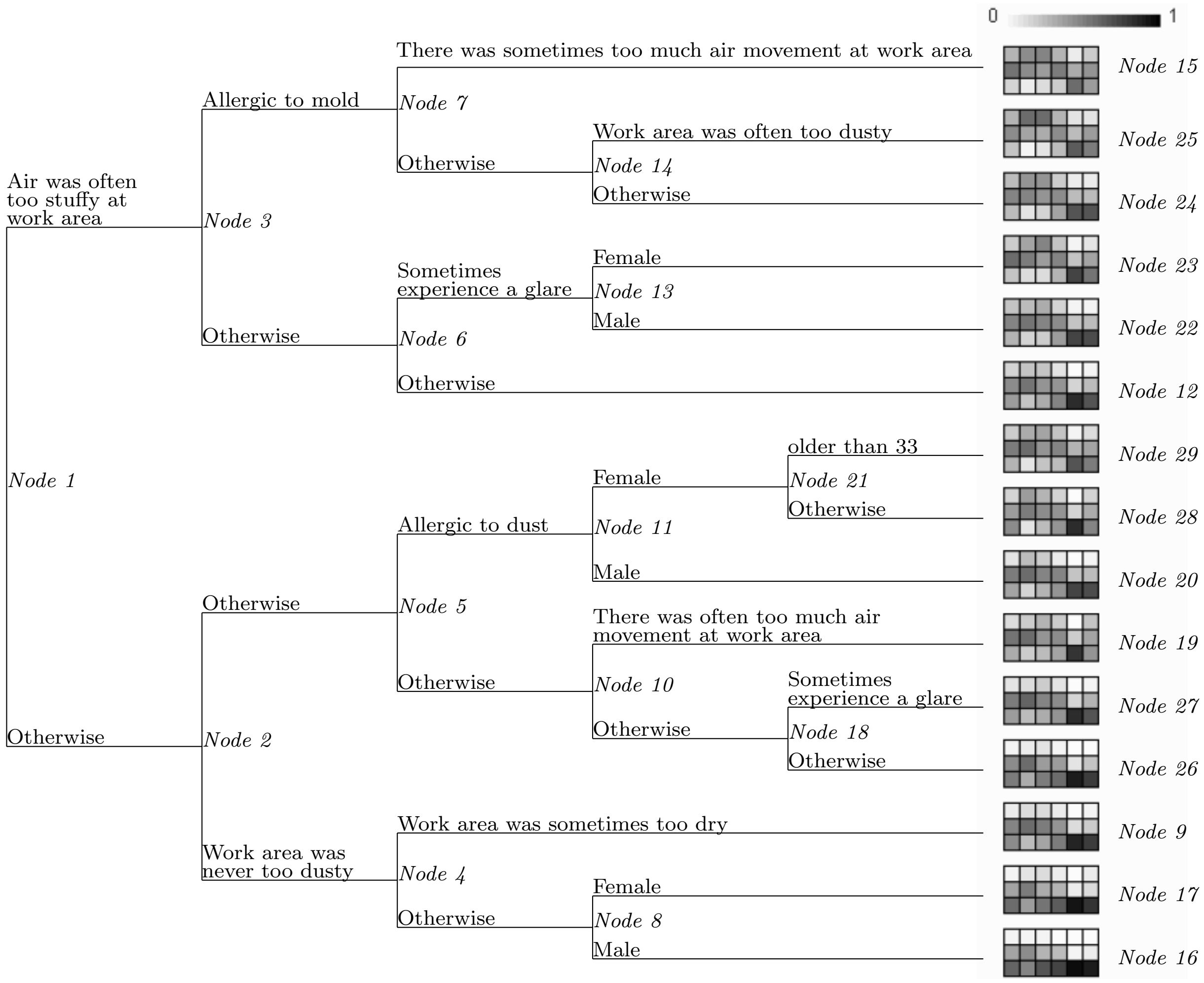


*Figure 4. Tree Structure for the Risk Factors of BROCS. Each Terminal Node is Represented by a $3 \times 6$ Table in which Rows Correspond to Levels of Response 0, 1, 2 from Bottom Up and Columns to the 6 Multivariate Responses. From Left to Right, They Are CNS, UA, Pain, Flu-like, Eyes, and LA. Each Cell Is Colored by Its Corresponding Frequency with the Color Scale Displayed on Top of Node 27.*

Dealing with multivariate responses, binary or ordinal, is a challenging and complicated task. Compared to the modeling of multiple binary responses, the task is even more complex for multiple ordinal responses. In an effort to reduce the computational burden, we proposed and evaluated a time-saving algorithm for optimization. A continuing effort to improve the computational efficiency remains worthwhile.

Through the analysis of a specific data set, we see that there are some clear advantages of analyzing ordinal responses. By beginning with the ordinal responses, we may also find helpful information to collapse the responses if we want to simplify the data and analysis.

Many other topics are worthwhile for future investigations. In this work, we focused on establishing an adaptive framework. In addition to the important issues mentioned above, with a large number of predictors, there may be alternative and competitive tree structures. To learn and utilize as much information as possible from such a data set, it is useful to consider forests such as bagging (Breiman 1996) and deterministic procedures (Zhang et al. 2003). Finally, Bayesian approach is also worth pursuing.

## APPENDIX: COMPUTATIONAL ISSUES

It follows from the node specific likelihood function (2) that

$$\begin{pmatrix} \frac{\partial}{\partial \psi_i} \\ \frac{\partial}{\partial \theta_i} \end{pmatrix} \log f_i = \begin{pmatrix} y_i - E(Y_i) \\ w_i - E(W_i) \end{pmatrix}.$$

Using the link matrices defined in (4) and based on model (5), we have

*Table 4. Comparison of Parameter Estimates Between the Simplied Method and the Quasi-Newton Algorithm when the Response is 3-Dimensional (1000 Replications)*

| | True Value | N=300 | | | | N=400 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Simplified Method | | Quasi-Newton | | Simplified Method | | Quasi-Newton | |
| | | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| $\psi_{11}$ | $-.75$ | $-0.75$ | 0.20 | $-0.75$ | 0.20 | $-0.75$ | 0.17 | $-0.75$ | 0.17 |
| $\psi_{12}$ | $-1.$ | $-1.02$ | 0.29 | $-1.02$ | 0.29 | $-1.02$ | 0.25 | $-1.02$ | 0.25 |
| $\psi_{21}$ | $-.75$ | $-0.75$ | 0.19 | $-0.75$ | 0.19 | $-0.75$ | 0.17 | $-0.75$ | 0.17 |
| $\psi_{22}$ | $-1.$ | $-1.00$ | 0.24 | $-1.00$ | 0.24 | $-1.01$ | 0.20 | $-1.01$ | 0.20 |
| $\psi_{31}$ | $-.75$ | $-0.75$ | 0.19 | $-0.75$ | 0.19 | $-0.74$ | 0.16 | $-0.74$ | 0.17 |
| $\psi_{32}$ | $-1.$ | $-1.00$ | 0.29 | $-1.00$ | 0.29 | $-1.01$ | 0.25 | $-1.00$ | 0.25 |
| $\theta_{11}$ | 0.25 | 0.24 | 0.16 | 0.24 | 0.16 | 0.24 | 0.16 | 0.24 | 0.15 |
| $\theta_{12}$ | 0.25 | 0.25 | 0.22 | 0.25 | 0.22 | 0.25 | 0.20 | 0.25 | 0.20 |
| $\theta_{21}$ | 0.25 | 0.26 | 0.23 | 0.26 | 0.23 | 0.26 | 0.20 | 0.26 | 0.20 |
| $\theta_{22}$ | 0.25 | 0.24 | 0.25 | 0.24 | 0.25 | 0.24 | 0.22 | 0.24 | 0.22 |

$$\begin{pmatrix} \frac{\partial}{\partial\psi} \\ \frac{\partial}{\partial\theta} \end{pmatrix} \log f_i = \begin{pmatrix} \frac{\partial\psi_i}{\partial\psi} & 0 \\ 0 & \frac{\partial\theta_i}{\partial\theta} \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial\psi_i} \\ \frac{\partial}{\partial\theta_i} \end{pmatrix} \log f_i$$
$$= \begin{pmatrix} B & 0 \\ 0 & C^T \end{pmatrix} \begin{pmatrix} y_i - E(Y_i) \\ w_i - E(W_i) \end{pmatrix},$$

where $B = I_{n\times(K-1)}$ and $C$ is defined in (4).

The likelihood function from the $N$ units of observations is

$$f(z, \psi, \theta) = \prod_{i=1}^{N} f_i(z_i, \psi, \theta),$$

and the log-likelihood is

$$l(z, \psi, \theta) = \sum_{i=1}^{N} l_i(z_i, \psi, \theta) = \sum_{i=1}^{N} \log f_i(z_i, \psi, \theta).$$

Thus, the score equations are

$$\begin{pmatrix} \frac{\partial}{\partial\psi} \\ \frac{\partial}{\partial\theta} \end{pmatrix} \log f = \sum_{i=1}^{N} \begin{pmatrix} B & 0 \\ 0 & C^T \end{pmatrix} \begin{pmatrix} y_i - E(Y_i) \\ w_i - E(W_i) \end{pmatrix}.$$

To find $\hat{\psi}$ and $\hat{\theta}$ that maximize the log-likelihood $l(z, \psi, \theta)$, we tested Newton's method, Quasi-Newton method and our simplified iteration procedure. The updating formula for Newton's method is listed below

(12)
$$\begin{pmatrix} \psi \\ \theta \end{pmatrix}^{(J+1)} = \begin{pmatrix} \psi \\ \theta \end{pmatrix}^{(J)} + (V^{(J)})^{-1} \sum_{i=1}^{N} \begin{pmatrix} B & 0 \\ 0 & C^T \end{pmatrix} \begin{pmatrix} y_i - E(Y_i) \\ w_i - E(W_i) \end{pmatrix},$$

where $V^{(J)}$ is the covariance matrix of (Y, W) at the J-th step. As noted in Zhang (1998), however, the computation for updating the estimators, especially the computation of $(V^{(J)})^{-1}$ can be very time consuming. Also as the number of repeated response $n$ increases, the method rarely converges. For example, when $n = 6$, the Newton's method rarely converged.

The BFGS method is one of the Quasi-Newton methods that avoids computing the inverse Hessian matrix. It calculates the score function and approximates the inverse Hessian matrix at step $k$. Let $\nu = \{\psi, \theta\}$. We first choose a initial guess $\nu_0$ and set $B_0 = I$, the identity matrix. At the $k+1$-th step of BFGS method, we proceed as follows:

1. Compute the score function $s_{k+1}$. If $s_{k+1}$=0, stop. Otherwise, let $d_{k+1} = B_k s_{k+1}$ ($B_k$ is the approximation of the inverse Hessian matrix).
2. Perform a linear search to find the optimal step length $\alpha_{k+1}$ and set $\nu_{k+1} = \nu_k + \alpha_{k+1} s_{k+1}$.
3. Update $B_{k+1}$, the approximation of inverse Hessian matrix. Let $v_k = s_{k+1} - s_k$ and $w_k = \nu_{k+1} - \nu_k$, then

$$B_{k+1} = B_k + \left(1 + \frac{v_k' B_k v_k}{v_k' w_k}\right) \frac{w_k w_k'}{v_k' w_k} - \frac{B_k v_k w_k' + w_k v_k' B_k}{v_k' w_k}.$$

The computation for BFGS method is also very extensive. It includes several vector and matrix operations during each updating step.

We adopt and verify a simplified iteration procedure to reduce the computational burden. Specifically, we use the sample variance-covariance matrix $V_0$ of

$$\begin{pmatrix} B & 0 \\ 0 & C^T \end{pmatrix} \begin{pmatrix} y_i \\ w_i \end{pmatrix}$$

and update the estimators of the parameters $\psi$ and $\theta$ as follows:

(13)
$$\begin{pmatrix} \psi \\ \theta \end{pmatrix}^{(J+1)} = \begin{pmatrix} \psi \\ \theta \end{pmatrix}^{(J)} + V_0^{-1} \sum_{i=1}^{N} \begin{pmatrix} B & 0 \\ 0 & C^T \end{pmatrix} \begin{pmatrix} y_i - E(Y_i) \\ w_i - E(W_i) \end{pmatrix}.$$

*Table 5. Comparison of Parameter Estimates Between the Simplied Method and the Quasi-Newton Algorithm when the Response is 4-Dimensional (1000 Replications)*

| | True Value | N=300 | | | | N=400 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Simplified Method | | Quasi-Newton | | Simplified Method | | Quasi-Newton | |
| | | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| $\psi_{11}$ | $-.75$ | $-0.74$ | 0.22 | $-0.74$ | 0.22 | $-0.75$ | 0.20 | $-0.75$ | 0.20 |
| $\psi_{12}$ | $-1.$ | $-1.01$ | 0.30 | $-1.01$ | 0.30 | $-1.00$ | 0.27 | $-1.00$ | 0.26 |
| $\psi_{21}$ | $-.75$ | $-0.75$ | 0.21 | $-0.75$ | 0.21 | $-0.75$ | 0.18 | $-0.75$ | 0.18 |
| $\psi_{22}$ | $-1.$ | $-1.01$ | 0.24 | $-1.01$ | 0.24 | $-1.01$ | 0.21 | $-1.00$ | 0.21 |
| $\psi_{31}$ | $-.75$ | $-0.74$ | 0.20 | $-0.74$ | 0.20 | $-0.74$ | 0.18 | $-0.74$ | 0.18 |
| $\psi_{32}$ | $-1.$ | $-1.03$ | 0.25 | $-1.03$ | 0.25 | $-1.02$ | 0.22 | $-1.01$ | 0.22 |
| $\psi_{41}$ | $-.75$ | $-0.74$ | 0.21 | $-0.74$ | 0.21 | $-0.74$ | 0.19 | $-0.74$ | 0.19 |
| $\psi_{42}$ | $-1.$ | $-1.01$ | 0.29 | $-1.01$ | 0.29 | $-1.00$ | 0.25 | $-1.00$ | 0.25 |
| $\theta_{11}$ | 0.25 | 0.24 | 0.12 | 0.24 | 0.12 | 0.24 | 0.10 | 0.24 | 0.10 |
| $\theta_{12}$ | 0.25 | 0.26 | 0.14 | 0.26 | 0.14 | 0.25 | 0.12 | 0.25 | 0.12 |
| $\theta_{21}$ | 0.25 | 0.25 | 0.15 | 0.25 | 0.14 | 0.25 | 0.13 | 0.25 | 0.13 |
| $\theta_{22}$ | 0.25 | 0.25 | 0.14 | 0.25 | 0.14 | 0.25 | 0.12 | 0.25 | 0.12 |

*Table 6. Comparison of Parameter Estimates Between the Simplied Method and the Quasi-Newton Algorithm when the Response is 5-Dimensional (1000 Replications)*

| | True Value | N=300 | | | | N=400 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Simplified Method | | Quasi-Newton | | Simplified Method | | Quasi-Newton | |
| | | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| $\psi_{11}$ | $-.75$ | $-0.75$ | 0.28 | $-0.75$ | 0.25 | $-0.75$ | 0.25 | $-0.75$ | 0.22 |
| $\psi_{12}$ | $-1.$ | $-1.00$ | 0.33 | $-0.99$ | 0.30 | $-1.01$ | 0.28 | $-1.00$ | 0.25 |
| $\psi_{21}$ | $-.75$ | $-0.74$ | 0.27 | $-0.75$ | 0.25 | $-0.74$ | 0.24 | $-0.74$ | 0.21 |
| $\psi_{22}$ | $-1.$ | $-1.01$ | 0.26 | $-1.00$ | 0.23 | $-1.01$ | 0.21 | $-1.00$ | 0.20 |
| $\psi_{31}$ | $-.75$ | $-0.74$ | 0.27 | $-0.74$ | 0.25 | $-0.74$ | 0.23 | $-0.74$ | 0.21 |
| $\psi_{32}$ | $-1.$ | $-1.02$ | 0.25 | $-1.00$ | 0.23 | $-1.02$ | 0.21 | $-1.00$ | 0.19 |
| $\psi_{41}$ | $-.75$ | $-0.75$ | 0.26 | $-0.75$ | 0.24 | $-0.75$ | 0.23 | $-0.75$ | 0.20 |
| $\psi_{42}$ | $-1.$ | $-1.01$ | 0.27 | $-0.99$ | 0.24 | $-1.01$ | 0.24 | $-0.99$ | 0.21 |
| $\psi_{51}$ | $-.75$ | $-0.74$ | 0.27 | $-0.74$ | 0.25 | $-0.74$ | 0.24 | $-0.74$ | 0.22 |
| $\psi_{52}$ | $-1.$ | $-1.02$ | 0.34 | $-1.00$ | 0.31 | $-1.01$ | 0.29 | $-1.00$ | 0.26 |
| $\theta_{11}$ | 0.25 | 0.24 | 0.11 | 0.25 | 0.09 | 0.24 | 0.10 | 0.25 | 0.08 |
| $\theta_{12}$ | 0.25 | 0.26 | 0.11 | 0.25 | 0.10 | 0.26 | 0.10 | 0.25 | 0.09 |
| $\theta_{21}$ | 0.25 | 0.26 | 0.12 | 0.25 | 0.10 | 0.26 | 0.10 | 0.25 | 0.09 |
| $\theta_{22}$ | 0.25 | 0.24 | 0.09 | 0.25 | 0.09 | 0.24 | 0.08 | 0.25 | 0.07 |

We report our simulation study to validate our simplified maximization procedure. Using model (5), we generate $N$ units of observations, $\{Y_i = (Y_{i1}, \ldots, Y_{in})^T\}_{i=1}^{N}$, where $N$ is chosen from 100 and 200. We consider the following four settings for $n$: (a) $n$=3; (b) $n$=4; (c) $n$=5; and (d) $n$=6. The level $K$ for the ordinal response is set to 3. The parameters to be estimated are $\psi_{11}$, $\psi_{12}, \ldots, \psi_{(n-1),1}$, $\psi_{(n-1),2}$, $\theta_{11}$, $\theta_{12}$, $\theta_{21}$, $\theta_{22}$.

Tables 4, 5, 6, and 7 report parameter estimates and their standard errors for the four settings of $n$, using our simplified method and the Quasi-Newton method. The results show that the simplified estimation procedure performs nearly identically to the Quasi-Newton method in all settings. It is also interesting to note that when $n = 6$, the Quasi-Newton method significantly underestimates the $\psi$ while our proposed simplified procedure has sufficient accuracy. As expected, an increased sample size improves the accuracy of the parameter estimates. Table 8 displays the dramatic savings in terms of computational time. In summary, Tables 4, 5, 6, 7, and 8 demonstrate that the simplified algorithm reduces computation time by 20 to 160 folds while maintaining the same or better level of accuracy as the Quasi-Newton method. This saving of computational time is significant in the present context because the selection of splits and the determination of the tree require numerous rounds of maximizing the likelihoods.

Table 7. Comparison of Parameter Estimates Between the Simplied Method and the Quasi-Newton Algorithm when the Response is 6-Dimensional (1000 Replications)

| | True | N=300 | | | | N=400 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Value | Simplified Method | | Quasi-Newton | | Simplified Method | | Quasi-Newton | |
| | | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. | Estimate | S.E. |
| $\psi_{11}$ | $-.75$ | $-0.72$ | 0.50 | $-0.62$ | 0.34 | $-0.72$ | 0.43 | $-0.60$ | 0.30 |
| $\psi_{12}$ | $-1.$ | $-0.99$ | 0.47 | $-0.99$ | 0.35 | $-1.00$ | 0.40 | 0.98 | 0.30 |
| $\psi_{21}$ | $-.75$ | $-0.70$ | 0.47 | $-0.60$ | 0.31 | $-0.70$ | 0.41 | $-0.59$ | 0.27 |
| $\psi_{22}$ | $-1.$ | $-1.00$ | 0.38 | $-0.99$ | 0.28 | $-1.00$ | 0.33 | 0.99 | 0.24 |
| $\psi_{31}$ | $-.75$ | $-0.71$ | 0.47 | $-0.61$ | 0.31 | $-0.71$ | 0.40 | $-0.60$ | 0.26 |
| $\psi_{32}$ | $-1.$ | $-1.00$ | 0.33 | $-0.98$ | 0.24 | $-1.00$ | 0.28 | $-0.98$ | 0.21 |
| $\psi_{41}$ | $-.75$ | $-0.71$ | 0.46 | $-0.61$ | 0.30 | $-0.71$ | 0.39 | $-0.60$ | 0.26 |
| $\psi_{42}$ | $-1.$ | $-1.01$ | 0.33 | $-0.99$ | 0.24 | $-1.01$ | 0.28 | $-0.99$ | 0.21 |
| $\psi_{51}$ | $-.75$ | $-0.69$ | 0.47 | $-0.59$ | 0.32 | $-0.70$ | 0.41 | $-0.60$ | 0.27 |
| $\psi_{52}$ | $-1.$ | $-1.01$ | 0.38 | $-0.99$ | 0.28 | $-1.00$ | 0.32 | $-0.98$ | 0.24 |
| $\psi_{61}$ | $-.75$ | $-0.70$ | 0.49 | $-0.60$ | 0.33 | $-0.70$ | 0.42 | $-0.59$ | 0.29 |
| $\psi_{62}$ | $-1.$ | $-1.02$ | 0.46 | $-1.01$ | 0.36 | $-1.02$ | 0.39 | $-1.00$ | 0.30 |
| $\theta_{11}$ | 0.25 | 0.24 | 0.14 | 0.22 | 0.07 | 0.24 | 0.12 | 0.21 | 0.06 |
| $\theta_{12}$ | 0.25 | 0.26 | 0.12 | 0.26 | 0.09 | 0.26 | 0.10 | 0.25 | 0.08 |
| $\theta_{21}$ | 0.25 | 0.25 | 0.12 | 0.25 | 0.09 | 0.25 | 0.10 | 0.25 | 0.08 |
| $\theta_{22}$ | 0.25 | 0.25 | 0.07 | 0.25 | 0.07 | 0.25 | 0.06 | 0.25 | 0.06 |

Table 8. Comparison of Running Time Between the Simplified Method and the Quasi-Newton Method (in Seconds)

| | N=300 | | | | N=400 | | | |
|---|---|---|---|---|---|---|---|---|
| $n_i$ | 3 | 4 | 5 | 6 | 3 | 4 | 5 | 6 |
| Simplified Method | 6 | 13 | 30 | 759 | 7 | 14 | 29 | 293 |
| Quasi-Newton | 387 | 1368 | 4100 | 16680 | 480 | 1699 | 4705 | 17978 |

*Received 10 October 2007*

## REFERENCES

Agresti, A. (1999). Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine* **18** 2191–2207.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Tress*. Pacific Grove, CA, Wadsworth.

Esposito, F., Malerba, D. and Semeraro, G. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**(5) 476–491.

Hedeker, D. and Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50** 933–944.

Kauermann, G. (2000). Modeling longitudinal data with ordinal response by varying coefficients. *Biometrics* **56** 692–698.

Li, X., Sweigart, J., Teng, J., Donohue, J. and Thombs, L. (2001). A dynamic programming based pruning method for decision trees. *Journal on Computing* **13** 332–345.

Mingers, J. (1987). Expert systems – rule induction with statistical data. *J. Operational Research Society* **28** 39–47.

Niblett, T. and Bratko, I. (1986). *Learning Decision Rules in Noisy Domains, Proc. Expert Systems 86*. Cambridge, Cambridge University Press.

Quinlan, J. R. (1987) Simplifying decision tree. *Machine Learning* **1** 221–234.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association* **87** 407–418.

Ten Have, T. R. (1996). A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* **52** 473–491. MR1395000

Zhang, H. P. (1997). Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics* **6** 74–91. MR1451991

Zhang, H. P. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association* **93**(441) 180–193.

Zhang, H. P., Feng, R. and Zhu, H. T. (2003). A latent variable model of segregation analysis for ordinal traits. *Journal of the American Statistical Association* **98**(464) 1023–1034. MR2041490

Zhang, H. P. and Singer, B. (1999). *Recursive Partitioning in the Health Science*. New York, Springer. MR1683316

Zhang, H. P., Wang, X. and Ye, Y. (2006). Detection of genes for ordinal traits in nuclear families and a unified approach for association studies. *Genetics* **172** 693–699.

Heping Zhang
Department of Epidemiology and Public Health
Yale University School of Medicine
New Haven, CT 06520-8034
E-mail address: heping.zhang@yale.edu

Yuanqing Ye
Department of Epidemiology
MD Anderson Cancer Center
Houston, TX 77030