Journal of Machine Learning Research 10 (2009) 141-158 Submitted 7/07; Revised 7/08; Published 2/09

Subgroup Analysis via Recursive Partitioning

Xiaogang Su XSU@PEGASUS.CC.UCF.EDU
Department of Statistics and Actuarial Science
University of Central Florida
Orlando, FL 32816, USA

Chih-Ling Tsai CLTSAI@UCDAVIS.EDU
Graduate School of Management
University of California, Davis
Davis, CA 95616, USA

Hansheng Wang HANSHENG@GSM.PKU.EDU.CN
Guanghua School of Management
Peking University
Beijing 100871, P. R. China

David M. Nickerson NICKERSN@MAIL.UCF.EDU
Department of Statistics and Actuarial Science,
University of Central Florida
Orlando, FL 32816, USA

Bogong Li LI.TIMOTHY@BLS.GOV
Office of Survey Methods Research
U.S. Bureau of Labor Statistics
Washington, DC 20212, USA

Editor: Saharon Rosset

Abstract

Subgroup analysis is an integral part of comparative analysis where assessing the treatment effect on a response is of central interest. Its goal is to determine the heterogeneity of the treatment effect across subpopulations. In this paper, we adapt the idea of recursive partitioning and introduce an interaction tree (IT) procedure to conduct subgroup analysis. The IT procedure automatically facilitates a number of objectively defined subgroups, in some of which the treatment effect is found prominent while in others the treatment has a negligible or even negative effect. The standard CART (Breiman et al., 1984) methodology is inherited to construct the tree structure. Also, in order to extract factors that contribute to the heterogeneity of the treatment effect, variable importance measure is made available via random forests of the interaction trees. Both simulated experiments and analysis of census wage data are presented for illustration.

Keywords: CART, interaction, subgroup analysis, random forests

1. Introduction

In comparative studies where two or more treatments are compared, subgroup analysis emerges after the overall assessment of the treatment effect and plays an important role in determining whether and how the treatment effect (i.e., comparison among treatments) varies across subgroups induced by covariates. In designed clinical trials, for example, practitioners and regulatory agencies are keen

©2009 Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson and Bogong Li.

to know if there are subgroups of trial participants who are more (or less) likely to be helped (or harmed) by the intervention under investigation. Subgroup analysis helps explore the heterogeneity of the treatment effect and extract the maximum amount of information from the available data. According to a survey conducted by Assmann et al. (2000), 70% of trials published over a three-month period in four leading medical journals involved subgroup analyses.

The research questions in subgroup analysis can be either pre-planned or raised in a post-hoc manner. If there is a prior hypothesis of the treatment effect being different in particular subgroups, then this hypothesis and its assessment should be part of the planned study design (CPMP, 1995). At the same time, subgroup analysis is also often used for post-hoc exploratory identification of unusual or unexpected results (Chow and Liu, 2004). Suppose, for example, that it is of interest to investigate whether the treatment effect is consistent among three age groups: young, middle-aged, and older individuals. The evaluation is formally approached by means of an interaction test between treatment and age. If the resulting test is significant, multiple comparisons are then used to find out further details about the magnitude and direction of the treatment effect within each age group. To ensure a valid experimentwise false positive rate, adjustment methods such as Bonferroni typed correction are often applied.

Limitations of traditional subgroup analysis have been extensively noted (see, e.g., Assmann et al., 2000, Sleight, 2000, and Lagakos, 2006). First of all, the subgroups themselves, as well as the number of subgroups to be examined, are specified by the investigator beforehand in the current practice of subgroup analysis, which renders subgroup analysis a highly subjective process. Even for the field expert, it is a daunting task to determine which specific subgroups should be used in subgroup analysis. The subjectivity may lead directly to dubious results and willful manipulations. For example, one may fail to identify a subgroup of great prospective interest or intentionally avoid reporting subgroups where the investigational treatment is found unsuccessful or even potentially harmful. Reliance on such analyses is likely to be erroneous and harmful. Moreover, significance testing is the main approach in subgroup analysis. Because there is no general guideline for selecting the number of subgroups, one has to examine numerous plausible possibilities to have a thorough assessment of the treatment effect. However, a large number of subgroups inevitably causes concerns related to multiplicity and lack of power.

In this paper, we propose a data-driven tree procedure, labelled as "interaction trees" (IT), to explore the heterogeneity structure of the treatment effect across a number of subgroups that are objectively defined in a post hoc manner. The tree method, also called recursive partitioning, was first proposed by Morgan and Sonquist (1963). By recursively bisecting the predictor space, the hierarchical tree structure partitions the data into meaningful groups and makes available a piecewise approximation of the underlying relationship between the response and its associated predictors. The applications of tree models have been greatly advanced in various fields especially since the development of CART (classification and regression trees) by Breiman et al. (1984). Their pruning idea for tree size selection has become and remains the current standard in constructing tree models.

The remainder of the paper is organized as follows. In Section 2, the IT procedure is presented in detail. Section 3 contains simulation studies designed for assessing the proposed method. In Section 4, we apply the IT procedure to analyze a wage data set, in which the goal is to determine whether or not women are underpaid or overpaid as compared to their male counterparts and, if so, by what amount. Section 5 concludes the paper with a brief discussion.

2. Tree-Structured Subgroup Analysis

We consider a study designed to assess a binary treatment effect on a continuous response or output while adjusting or controlling for a number of covariates. Suppose that the data available contain n i.i.d. observations $\{(y_i, \mathrm{trt}_i, \mathbf{x}_i) : i = 1, \ldots, n\}$, where y_i is the continuous response; trt_i is the binary treatment indicator taking values 1 or 0; and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is the associated p-dimensional covariate vector where its components can be of mixed types, that is, categorical or continuous. Our goal in subgroup analysis is to find out whether there exist subgroups of individuals in which the treatment shows heterogeneous effects, and if so, how the treatment effect varies across them.

Applying a tree procedure to guide the subgroup analysis is rather intuitive. First, subgroup analysis essentially involves the interaction between the treatment and the covariates. The tree algorithm is known as an excellent tool for exploring interactions. As a matter of fact, the first implementation of decision trees is referred to as Automatic Interaction Detection (AID; Morgan and Sonquist, 1963). However, tree methods handle interactions *implicitly*. It is often hard to determine whether or not interaction really exists among variables for a given tree structure. We shall seeks ways that enable us to *explicitly* assess the interaction between the treatment and the covariates. Secondly, the hierarchical binary tree structure naturally groups data in an optimal way. By recursively partitioning the data into two subgroups that show the greatest heterogeneity in treatment effect, we are able to optimize the subgroup analysis and make it more efficient in representing the heterogeneity structure of the treatment effect. Thirdly, the tree procedure is objective, data-driven, and automated. The grouping strategy and the number of subgroups are automatically determined by the procedure. The proposed method would result in a set of objectively recognized and mutually exclusive subgroups, ranking from the most effective to the least effective in terms of treatment effect. To the best of our knowledge, Negassa et al. (2005) and Su et al. (2008), who studied tree-structured subgroup analysis in the context of censored survival data, are the only previous works along similar lines to our proposal.

To construct the IT model, we follow the CART (Breiman et al., 1984) convention, which consists of three major steps: (1) growing a large initial tree; (2) a pruning algorithm; and (3) a validation method for determining the best tree size. Once a final tree structure is obtained, the subgroups are naturally induced by its terminal nodes. To achieve better efficiency, an amalgamation algorithm is used to merge terminal nodes that show homogenous treatment effects. In addition, we adopt the variable importance technique in the context of random forest to extract covariates that exhibit important interactions with the treatment.

2.1 Growing a Large Initial Tree

We start with a single split, say s, of the data. This split is induced by a threshold on a predictor X. If X is continuous, then the binary question whether $X \leq c$ is considered. Observations answering "yes" go to the left child node t_L and observations answering "no" go to the right child node t_R. If X is nominal with r distinct categories $C = \{c_1, \ldots, c_r\}$, then the binary question becomes "Is $X \in A$?" for any subset $A \subset C$. When X has many distinct categories, one may place them into order according to the treatment effect estimate within each category and then proceed as if X is ordinal. This strategy helps reduce the computational burden. A theoretical justification is provided by Theorem 1 in the appendix.

For a given node t, a split s yields the following 2×2 table:

	child node							
treatment	t_L				t_R			
1	μ_1^L	$\bar{y}_1^L$	s_1^2	n_1	μ_1^R	$\bar{y}_1^R$	s_2^2	n_2
0	μ_0^L	$\bar{y}_0^L$	s_3^2	n_3	μ_0^R	$\bar{y}_0^R$	s_4^2	n_4

Here, $\{\mu_1^L, \bar{y}_1^L, s_1^2, n_1\}$ are the population mean, the sample mean, the sample variance, and the sample size for the treatment 1 group in the left child node t_L, respectively. Similar notation applies to the other quantities. To evaluate the heterogeneity of the treatment effect between t_L and t_R, we compare $(\mu_1^L - \mu_0^L)$ with $(\mu_1^R - \mu_0^R)$ so that the interaction between split s and treatment is under investigation.

A natural measure for assessing the interaction is given by

$$t(s) = \frac{(\bar{y}_1^L - \bar{y}_0^L) - (\bar{y}_1^R - \bar{y}_0^R)}{\hat{\sigma} \cdot \sqrt{1/n_1 + 1/n_2 + 1/n_3 + 1/n_4}}, \tag{1}$$

where $\hat{\sigma}^2 = \sum_{i=1}^4 w_i s_i^2$ is a pooled estimator of the constant variance, and $w_i = (n_i - 1)/\sum_{j=1}^4 (n_j - 1)$. For a given split s, $G(s) = t^2(s)$ converges to a $\chi^2(1)$ distribution. We will use $G(s)$ as the splitting statistic, while remaining aware that the sign of $t(s)$ supplies useful information regarding the direction of the comparison.

The best split $s^\star$ is the one that yields the maximum G statistic among all permissible splits. That is,

$$G(s^\star) = \max_s G(s).$$

The data in node t are then split according to the best split $s^\star$. The same procedure is applied to split both child nodes: t_L and t_R. Recursively doing so results in a large initial tree, denoted by T_0.

Remark 1: It can be easily seen that $t(s)$ given in (1) is equivalent to the t test for testing $H_0 : \gamma = 0$ in the following threshold model

$$y_i = \beta_0 + \beta_1 \cdot \mathrm{trt}_i + \delta \cdot z_i^{(s)} + \gamma \cdot \mathrm{trt}_i \cdot z_i^{(s)} + \varepsilon_i, \tag{2}$$

where $z_i^{(s)} = 1_{\{X_i \leq c\}}$ is the binary variable associated with split s. This observation sheds light on extensions of the proposed method to situations involving different types of responses and multi-level treatments. Note that inclusion of both trt_i and $z_i^{(s)}$ in model (2) is important to assure the correspondence between the regression coefficient γ and the interaction, that is, $\gamma = (\mu_1^L - \mu_0^L) - (\mu_1^R - \mu_0^R)$.

Remark 2: In constructing the large initial tree T_0, a terminal node is declared when any one of the following conditions is met: the node gets pure in the sense that all the covariates have the same values; the total number of observations in the node is less than some preset minimum node size; the depth of the node is greater than some preset maximum tree depth; for all permissible splits, the minimum of (n_1, n_2, n_3, n_4) is below some preset threshold. The last condition is particularly useful in preventing the "end-cut preference" problem as discussed in CART (Breiman et al., 1984, pp. 313–317). For this purpose, the threshold can be set adaptively with respect to the depth of the node, following the suggestion of Torgo (2001).

2.2 Pruning

The final tree could be any subtree of T_0. To narrow down the choices, Breiman et al. (1984) proposed a pruning algorithm which results in a sequence of optimally pruned subtrees by iteratively truncating the "weakest link" of T_0. In the following, we briefly describe an interaction-complexity pruning procedure, which is analogous to the split-complexity pruning algorithm in LeBlanc and Crowley (1993). We refer the reader to their paper for a detailed description.

To evaluate the performance of an interaction tree T, we define an interaction-complexity measure:

$$G_\lambda(T) = G(T) - \lambda \cdot \mid T - \widetilde{T} \mid, \tag{3}$$

where $G(T) = \sum_{h \in T - \widetilde{T}} G(h)$ measures the overall amount of interaction in T; the total number of internal nodes $|T - \widetilde{T}|$ corresponds to the complexity of the tree; and the complexity parameter $\lambda \geq 0$ acts as a penalty for each added split. Given a fixed λ, a tree with larger $G_\lambda(T)$ is preferable.

Start with T_0. For any internal node h of T_0, calculate $g(h) = G(T_h)/ \mid T_h - \widetilde{T_h} \mid$, where T_h is the branch with h as root and $|T_h - \widetilde{T_h}|$ denotes the number of internal nodes of T_h. Then, the node $h^\star$ with smallest $g(h^\star)$ is the "weakest link," in the sense that $h^\star$ becomes ineffective first as λ increases. Next, let T_1 be the subtree after pruning off the branch $T_{h^\star}$ from T_0, and subsequently apply the same procedure to prune T_1. Repeating this procedure yields a nested sequence of subtrees $T_M \prec \cdots \prec T_m \prec T_{m-1} \prec \cdots \prec T_1 \prec T_0$, where T_M is the null tree with the root node only and $\prec$ means "is a subtree of".

2.3 Selecting the Best-Sized Subtree

The final tree will be selected from the nested subtree sequence $\{T_m : m = 0, 1, \ldots, M\}$, again, based on the maximum interaction-complexity measure $G_\lambda(T_m)$ given in (3). For tree size determination purposes, λ is suggested to be fixed within the range $2 \leq \lambda \leq 4$, where $\lambda = 2$ is in the spirit of the Akaike information criterion (AIC; Akaike, 1973) and $\lambda = 4$ corresponds roughly to the 0.05 significance level on the $\chi^2(1)$ curve. Another choice is $\lambda = \ln(n)$, citing the Bayesian information criterion (BIC; Schwarz, 1978).

To overcome the over-optimism due to greedy search, an "honest" estimate of the goodness-of-split measure $G(T_m)$ is needed. This can be achieved by a validation method. If the sample size is large, $G(T_m)$ can be recalculated using an independent subset of the data (termed the validation sample). If the sample size is small or moderate, one has to resort to techniques such as v-fold cross-validation or bootstrapping methods in order to validate $G(T_m)$. The bootstrap method used in LeBlanc and Crowley (1993), for example, can be readily adapted to interaction trees.

Remark 3: As another alternative, one referee suggested the use of in-sample BIC to determine the optimally-sized subtree, which can be promising, although further research effort is needed to determine the effective degrees of freedom (see, e.g., Ye, 1998, and Tibshirani and Knight, 1999) associated with interaction trees.

2.4 Summarizing the Terminal Nodes

The total number of subgroups, which could be further reduced, corresponds to the automatically selected best tree size. Existence of the overall interaction between the treatment and the covariates can be roughly assessed by inspecting whether a null final tree structure is obtained. Unlike conventional subgroup analysis, the subgroups obtained from the IT procedure are mutually exclusive. Due to the subjectivity in determining the subgroups and multiplicity emerging from significance testings across subgroups, it is generally agreed that subgroup analysis should be regarded as exploratory and hypothesis-generating. As recommended by Lagakos (2006), it is best not to present p-values for within-subgroup comparisons, but rather to give an estimate of the magnitude of the treatment difference. To summarize the terminal nodes, one can compute the average responses for both treatments within each terminal node, as well as their relative differences or ratios and the associated standard errors. If, however, one would like to perform some formal statistical testings, then it is important to conduct these tests based on yet another independent sample. To do so, one partitions the data into three sets: the learning sample $\mathcal{L}_1$, the validation sample $\mathcal{L}_2$, and the test sample $\mathcal{L}_3$. The best IT structure will be developed using $\mathcal{L}_1$ and $\mathcal{L}_2,$ and then reconfirmed using the test sample $\mathcal{L}_3$.

It happens often that the treatment shows homogeneous effects for entirely different causal reasons in terminal nodes stemming from different branches. In this case, a merging scheme analogous to the approach of Ciampi et al. (1986) can be useful to further bring down the number of final subgroups. A smaller number of subgroups are easier to summarize and comprehend and more efficient to represent the heterogeneity structure for the treatment effect. It also ameliorates the multiplicity issue in within-subgroup comparisons. In this scheme, one computes the t statistic in equation (1) between every pair of the terminal nodes. The pair showing the least heterogeneity in treatment effect are then merged together. The same procedure is executed iteratively until all the remaining subgroups display outstanding heterogeneity. In the end, one can sort the final subgroups based on the strength of the treatment effect, from the most effective to the least effective.

Remark 4: The hierarchical tree structure is appealing mainly because of its easy interpretability. Apparently the merging scheme would invalidate the tree structure, yet lead to better interpretations. It is noteworthy that this merging scheme contributes additional optimism to the results. Thus, we suggest that amalgamation be executed with data $(\mathcal{L}_1 + \mathcal{L}_2)$ that pool the learning sample $\mathcal{L}_1$ and the validation sample $\mathcal{L}_2$ together, so that the resultant subgroups can be validated using the test sample $\mathcal{L}_3$. We shall illustrate this scheme with an example presented in Section 4.

2.5 Variable Importance Measure via Random Forests

Variable importance measure is another attractive feature offered by recursive partitioning. In the context of subgroup analysis, a covariate is called an *effect-modifier* of the treatment if it strongly interacts with the treatment. Variable importance measure helps answer questions such as which features or predictors are important in modifying the treatment effect. This issue cannot be fully addressed by simply examining the splitting variables shown in a single final IT structure, as an important variable can be completely masked by other correlated ones. While there are many methods available for extracting variable importance information, we propose an algorithm analogous to the procedure used in random forests (Breiman, 2001), which is among the newest and most promising developments in this regards.

Algorithm 1: Computing Variable Importance Measure via Random Forests.

```
Initialize all $V_j$'s to 0.
For $b = 1, 2, \ldots, B$, do
    • Generate bootstrap sample $\mathcal{L}_b$ and obtain the out-of-bag sample $\mathcal{L} - \mathcal{L}_b$.
    • Based on $\mathcal{L}_b$, grow a large IT tree $T_b$ by searching over $m_0$ randomly selected
    covariates at each split.
    • Send $\mathcal{L} - \mathcal{L}_b$ down $T_b$ to compute $G(T_b)$.
    • For all covariates $X_j$, $j = 1, \ldots, p$, do
        ∘ Permute the values of $X_j$ in $\mathcal{L} - \mathcal{L}_b$.
        ∘ Send the permuted $\mathcal{L} - \mathcal{L}_b$ down to $T_b$ to compute $G_j(T_b)$.
        ∘ Update $V_j \leftarrow V_j + \frac{G(T_b) - G_j(T_b)}{G(T_b)}$.
    • End do.
End do.
Average $V_j \leftarrow V_j / B$.
```

Let V_j denote the importance measure of the j-th covariate or feature X_j for $j = 1, \ldots, p$. We construct random forests of IT trees by taking B bootstrap samples $\mathcal{L}_b$, $b = 1, \ldots, B$. This is done by searching over only a subset of randomly selected m_0 covariates at each split. For each IT tree T_b, the b-th out-of-bag sample (denoted as $\mathcal{L} - \mathcal{L}_b$), which contains all observations that are not in $\mathcal{L}_b$, is sent down T_b to compute the interaction measure $G(T_b)$. Next, the values of the j-th covariate in $\mathcal{L} - \mathcal{L}_b$ are randomly permuted. The permuted out-of-bag sample is then sent down T_b to recompute $G_j(T_b)$. The relative difference between $G(T_b)$ and $G_j(T_b)$ is recorded. The procedure is repeated for B bootstrap samples. As a result, the importance measure V_j is the average of the relative differences over all B bootstrap samples. The whole procedure is summarized in Algorithm 1.

The variable importance technique in random forests has been increasingly studied in its own right and applied as a tool for variable selection in various fields. This method generally belongs to the "cost-of-exclusion" (Kononenko and Hong, 1997) feature selection category, in which the importance or relevance of a feature is determined by the difference in some model performance measure with and without the given feature included in the modeling process. In Algorithm 1, random forests make available a flexible modeling process while exclusion of a given feature is carried out by permutating its values.

3. Simulated Studies

This section contains simulated experiments designed to investigate the performance of the IT procedure. We generate data from six models outlined in Table 1. Each data set consists of a continuous response Y, a binary treatment, and four covariates X_1–X_4 simulated from a discrete uniform distribution over $(0.02, 0.04, \ldots, 1.00)$. However, only a subset of the covariates interact with the treatment.

Model	Form	Error Distribution
A	$Y = 2+2\cdot\text{trt}+2Z_1+2Z_2+\varepsilon$	$\mathcal{N}(0,1)$
B	$Y = 2+2\cdot\text{trt}+2Z_1+2Z_2+2\cdot\text{trt}\cdot Z_1 Z_2+\varepsilon$	$\mathcal{N}(0,1)$
C	$Y = 2+2\cdot\text{trt}+2Z_1+2Z_2+2\cdot\text{trt}\cdot Z_1+2\cdot\text{trt}\cdot Z_2+\varepsilon$	$\mathcal{N}(0,1)$
D	$Y = 10+10\cdot\text{trt}\cdot\exp\{(X_1-0.5)^2+(X_2-0.5)^2\}+\varepsilon$	$\mathcal{N}(0,1)$
E	$Y = 2+2\cdot\text{trt}+2Z_1+2Z_2+2\cdot\text{trt}\cdot Z_1+2\cdot\text{trt}\cdot Z_2+\varepsilon$	$\text{Unif}(-\sqrt{3},\sqrt{3})$
F	$Y = 2+2\cdot\text{trt}+2Z_1+2Z_2+2\cdot\text{trt}\cdot Z_1+2\cdot\text{trt}\cdot Z_2+\varepsilon$	$\exp(1)$

Table 1: Models used for assessing the performance of the IT procedure. Note that $Z_1 = 1_{\{X_1 \leq 0.5\}}$ and $Z_2 = 1_{\{X_2 \leq 0.5\}}$.

Specifically, Model A is a plain additive model with no interaction. It helps assess the type I error or false positive rate when using the IT procedure. Model B involves a second-order interaction between the treatment and two terms of thresholds, both at 0.5, on X_1 and X_2. If the IT procedure works, a tree structure with three terminal nodes should be selected. In Model C, the two thresholds, each interacting with the treatment, are present in an additive manner. Model D involves interaction of complex forms other than cross-products. A large tree is expected in order to represent the interaction structure in this case. Models E and F are similar to Model C, but with different error distributions. They are useful in evaluating the robustness of the IT procedure to deviations from normality.

Only one set of sample sizes is reported: 800 observations in the learning sample $\mathcal{L}_1$ and 400 observations in the validation sample $\mathcal{L}_2$. Each model is examined for 200 simulation runs and four choices of λ, $\{2,3,4,\ln(400)\}$, are considered for determining the best tree structure. The relative frequencies of the final tree sizes selected by the IT procedure are presented in Table 2. The expected final tree size for each model is highlighted in boldface. Note that both X_1 and X_2, but neither X_3 nor X_4, are actually involved in models B-F. To address the variable selection issue, we count the frequency of "hits" (i.e., the final tree selected by the IT procedure is split by X_1 and X_2 and only by them). The results are presented in the last column of Table 2.

We first examine the results for Model A, which involves no interaction. The IT procedure correctly selects the null tree structure at least 83.5% of the time. When $\lambda = \ln(n)$, the percentage of correct selections becomes 98.5%. The 98.5% of correct selections yields an empirical size of 100%-98.5% = 1.5%, which is well within the acceptable level. This implies that the chance for the IT procedure to extract an unsolicited interactions is really small. For models B-F, the IT procedure also successfully identifies the true final tree structure and selects the desired splitting variables a majority of the time. Moreover, from results for models E and F, it seems rather robust against deviations from normality. When comparing different complexity parameters, we see that $\lambda = \ln(n)$ provides the best selection. This is mainly because of the large sample size and relatively strong signals considered in our simulation configuration (see, e.g., McQuarrie and Tsai, 1998).

To evaluate the proposed variable importance technique, we generate a data set containing 1,200 observations from each model. Then a total number of 500 random interaction trees with $m_0 = 1$ are used to compute the variable importance. In Figure 1, graphs A-I, B-I, ..., F-I display the resultant importance scores for models A-F, respectively. For comparison, we also apply a simple feature

	Complexity	Final Tree Size							
Model	λ	1	2	3	4	5	6	≥ 7	Hits
A	2	**83.5**	5.0	2.5	2.5	2.0	0.5	4.0	83.5
	3	**94.0**	3.5	0.5	0.5	1.0	0.5	0.0	94.0
	4	**97.5**	2.5	0.0	0.0	0.0	0.0	0.0	97.5
	$\ln(n)$	**98.5**	1.5	0.0	0.0	0.0	0.0	0.0	98.5
B	2	0.0	0.0	**67.0**	9.0	9.0	2.5	12.5	77.5
	3	0.0	0.0	**83.0**	6.5	5.5	1.0	4.0	90.0
	4	0.0	0.0	**89.0**	6.5	3.0	1.0	0.5	95.0
	$\ln(n)$	1.0	0.0	**91.5**	6.5	2.0	0.0	0.0	97.5
C	2	0.0	0.0	0.0	**66.5**	10.5	9.5	13.5	77.5
	3	0.0	0.0	0.0	**82.5**	7.5	6.0	4.0	90.5
	4	0.0	0.0	0.0	**88.0**	6.5	4.5	1.0	95.0
	$\ln(n)$	0.0	0.0	0.0	**94.0**	4.0	1.5	0.5	98.5
D	2	0.0	0.0	0.0	0.0	13.0	10.5	76.5	66.5
	3	0.0	0.0	0.0	0.5	24.5	16.0	59.0	83.5
	4	0.0	0.0	0.0	0.5	35.5	18.5	45.5	91.5
	$\ln(n)$	0.0	0.0	2.0	3.5	54.0	16.0	24.5	96.0
E	2	0.0	0.0	0.0	**74.0**	12.0	6.0	8.0	82.0
	3	0.0	0.0	0.0	**88.5**	7.5	3.5	0.5	93.0
	4	0.0	0.0	0.0	**93.5**	5.0	1.0	0.5	97.5
	$\ln(n)$	0.0	0.0	0.0	**97.0**	2.5	0.0	0.5	98.5
F	2	0.0	0.0	0.0	**67.5**	13.0	8.0	11.5	76.5
	3	0.0	0.0	0.0	**84.5**	7.0	4.5	4.0	91.0
	4	0.0	0.0	0.0	**90.0**	7.5	1.5	1.0	96.5
	$\ln(n)$	0.0	0.0	0.0	**95.0**	4.5	0.5	0.0	99.0

Table 2: Relative frequencies (in percentages) of the final tree size identified by the interaction tree (IT) procedure in 200 runs.

selection approach, in which the importance of a covariate, say, X, is determined by the p-value for testing $H_0 : \beta_3 = 0$ in the interaction model $y = \beta_0 + \beta_1 x + \beta_2 \,\mathrm{trt} + \beta_3 \,\mathrm{trt} \cdot x + \varepsilon$. In Figure 1, graphs A-II, B-II, ..., F-II depict the resulting logworths for models A-F, where the logworth is defined as minus base 10 logarithm of the p-value. Both methods are able to pick up important variables in many cases. Nevertheless, the latter approach is focused on cross-product and first-order interactions only, which accounts for its failure in models B and D. In Figure B-II, it fails to identify X_1

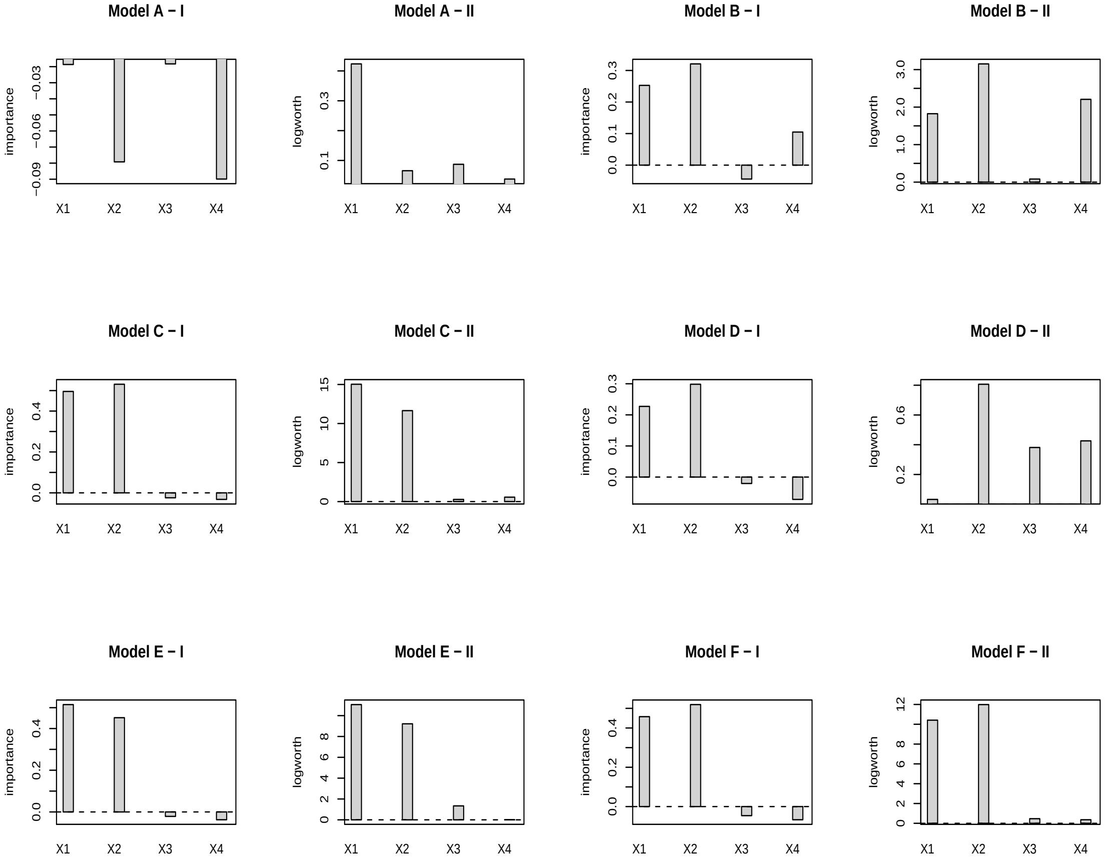

Figure 1: Variable Importance via Random Forests and Feature Selection for the Six Simulation Models. With a data set consisting of 1200 observations generated from Model A, Graph A-I plots the variable importance score computed from 500 random interaction trees while Graph A-II plots the logworth of the p-value from the simple feature selection method. Note that logworth $= -\log_{10}$(p-value). Similar methods were used to make other graphs.

and mistakenly selects X_4. In Figure D-II, none of the covariates are found important.

Remark 5: Comparatively, the random forest approach facilitates better comprehensive evaluation by automatically taking into consideration interactions of higher orders and complex forms. As pointed out by an anonymous reviewer, some other feature selection methods such as the wrapper and/or embedded algorithms can be adopted to detect interactions. It would be interesting in future research to have a more comprehensive study in comparing various algorithms with the variable importance measure obtained from random forests.

	Name	Type	Levels	Description
X_1	age	continuous	70	age
X_2	workclass	categorical	4	class of worker
X_3	education	ordinal	16	education levels
X_4	marital	categorical	7	marital status
X_5	industry[1]	categorical	22	major industry
X_6	occupation[2]	categorical	13	major occupation
X_7	race	categorical	5	race
X_8	union	categorical	2	member of a labor union
X_9	fulltime	categorical	6	full or part time employment
X_{10}	tax.status	categorical	6	tax filer status
X_{11}	household.sum[3]	categorical	7	detailed household summary
X_{12}	n.emplyee	ordinal	7	number of persons worked for employer
X_{13}	country.birth	categorical	39	country of birth
X_{14}	citizenship	categorical	7	US citizen or not
X_{15}	self.employed	categorical	3	own business or self employed
X_{16}	weeks.worked	continuous	53	weeks worked in year

[1] The specific levels for industry (X_5): 1 - Agriculture; 2 - Business and repair services; 3 - Communications; 4 - Construction; 5 - Education; 6 - Entertainment; 7 - Finance insurance and real estate; 8 - Forestry and fisheries; 9 - Hospital services; 10 - Manufacturing-durable goods; 11 - Manufacturing-nondurable goods; 12 - Medical except hospital; 13 - Mining; 14 - Other professional services; 15 - Personal services except private HH; 16 - Private household services; 17 - Public administration; 19 - Retail trade; Social services; 20 - Transportation; 21 - Utilities and sanitary services; 22 - Wholesale trade.

[2] The specific levels for occupation (X_6): 1- Adm support including clerical; 2 - Executive admin and managerial; 3 - Farming forestry and fishing; 4 - Handlers equip cleaners etc; 5 - Machine operators assmblrs & inspctrs; 6 - Other service; 7 - Precision production craft & repair; 8 - Private household services; 9 - Professional specialty; 10 - Protective services; 11 - Sales; 12 - Technicians and related support; 13 - Transportation and material moving.

[3] The specific levels for household.sum (X_{11}): 1 - Child 18 or older; 2 - Child under 18 never married; 3 - Group Quarters, Secondary individual; 4 - Householder; 5 - Nonrelative of householder; 6 - Other relative of householder; 7 - Spouse of householder.

Table 3: Variable Description for the Census Wage Data.

4. An Example - The CPS Data

Society has long been arguing for pay equality between women and men. Although the pay gap has narrowed, according to current statistics, gaps between the two sexes still exist. For example, the Bureau of Labor Statistics of the U.S. Department of Labor, in 2004, the most recent year for which statistics are available, reported that women's median weekly earnings were only 80 percent that of men. This represents an improvement over 1979, when women brought home only 62 percent of earnings compared to their male counterparts. A public policy advocate would be very interested in specific subgroups of the working population where the pay gap between sexes is still dominant. Traditional statistical methods consider simple cross tabulations according to a number of variables.

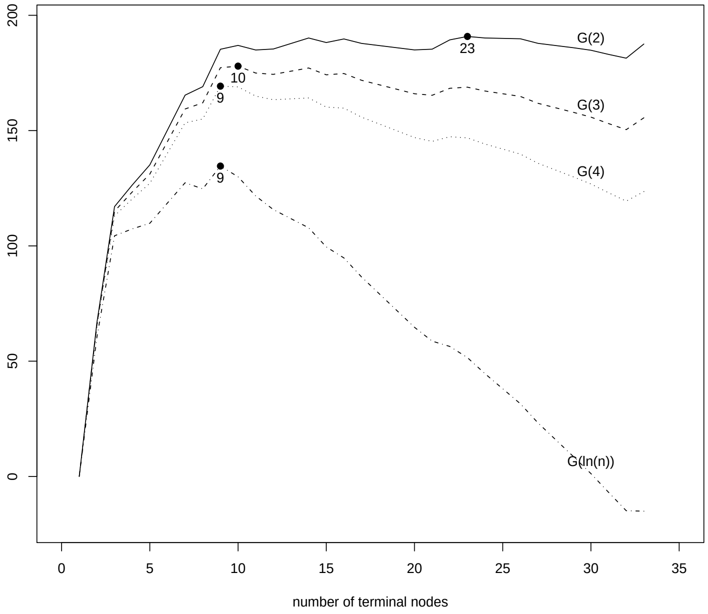

Figure 2: Tree size selection for the CPS data (the plot of $G_\lambda(T_m)$ vs. tree size).

The quest becomes complicated when there are a large number of classification variables to choose from, which motivates the IT procedure.

We extract data from the Current Population Survey (CPS) database, which can be accessed at (`http://www.census.gov/cps/`). The CPS is a national monthly survey of approximately 60,000 households conducted by the U.S. Census Bureau for the Bureau of Labor Statistics. Information collected in the survey includes employment status, hours worked and income from work as well as a number of demographic characteristics of the household members. From the CPS 1995 March Supplement, we compile a data set, which contains 16,602 individuals with no missing value involved. We use hourly wages in U.S. dollars as a measure of pay. Besides gender, there are a total of 16 demographical covariates included. A brief description for these covariates is given in Table 3. Several of them are nominal with many levels. We only list detailed codings for those variables appearing in the final tree structure.

To apply IT, we take a logarithmic transformation on wage. Then, we randomly divide the entire data (denoted as $\mathcal{L}$) into three sets with a ratio of approximately 2 : 1 : 1. A large initial tree with 33 terminal nodes is constructed and pruned using data in $\mathcal{L}_1$. Sending the validation sample $\mathcal{L}_2$ down each subtree, Figure 2 depicts the resultant $G_\lambda(T_m)$ score versus the subtree size. It can be seen that the four choices of complexity parameter $\lambda = 2, 3, 4$, and $\ln(n)$ yield the best tree sizes of 23, 10, 9, and 9, respectively. For the sake of illustration, the best tree structure with 9 terminal nodes as well as some related summary statistics are given in Figure 3. To merge those terminal nodes among which the wage discrepancies due to gender are not significantly different, we run the amalgamation algorithm with the pooled data $(\mathcal{L}_1 + \mathcal{L}_2)$. It results in four final subgroups, which are then ranked as I-IV according to the ratio of women versus men in terms of average wage. In Group

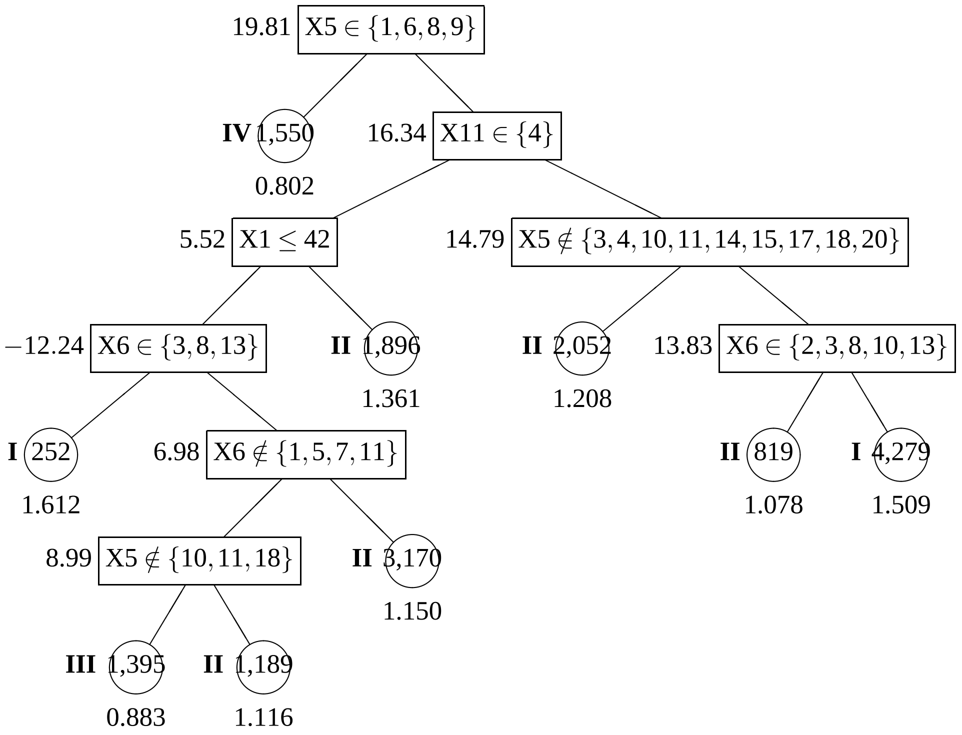

Figure 3: The best-sized interaction tree for the CPS data selected by BIC. For each internal node denoted by a box, the splitting rule is given inside the box. Observations satisfying the condition go to the left node while observations not satisfying the condition go the right node. To the left of each internal node is the $t(s)$ test statistic recomputed using $\mathcal{L}_3$. Terminal nodes are denoted by circles and ranked with Roman numerals on the left. The node size (i.e., number of individuals) is given inside the circle and the ratio of average wages for males versus females is given underneath, both based on the entire data set $\mathcal{L}$.

I, women are most underpaid compared to men. In Group II, women are somewhat underpaid, etc. This specification was also made available to each terminal node in Figure 3.

Table 4 summarizes the final four subgroups. For each subgroup, the number of men and women, as well as their average wages, are computed based on the whole data $\mathcal{L}$. The two-sample t test for comparing the average wages between men and women is also presented. This test was computed by using the test sample $\mathcal{L}_3$. Since there are four tests performed, one may compare the resultant p-values with $0.05/4 = 0.0125$ by applying the Bonferroni-typed adjustment to the joint significant level $\alpha = 0.05$. As a result, Groups I, II, and IV, each marked with an asterisk, show significant differences in wage between men and women.

Figure 3 indicates that the wage disparity between men and women varies with their occupation (X_5), industry of the job (X_6), household compositional situation (X_{11}), and age (X_1). For both Groups I and II, which constitute the majority of the population, women are paid significantly less than men. It is particularly pronounced in Group I, where the average wage of men is \$12.29 per

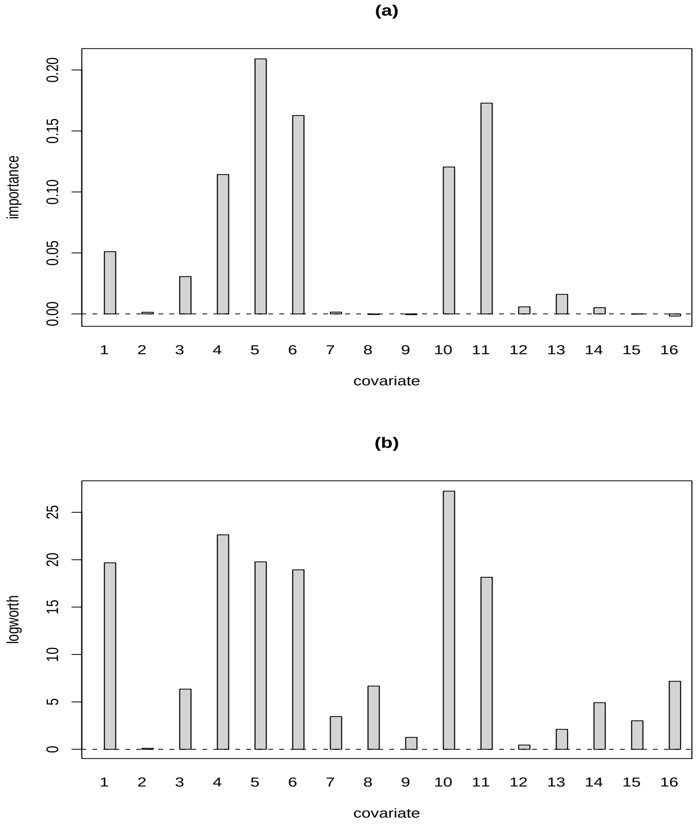

Figure 4: Variable importance measures for the CPS Data: (a), via random forests of interaction trees; (b), based on the p-value associated with the interaction terms in an interaction model that includes the treatment and the covariate only. Note that the logworth is $-\log_{10}(\text{p-value})$. The sixteen covariates are `age` (X_1), `workclass` (X_2), `education` (X_3), `marital` (X_4), `industry` (X_5), `occupation` (X_6), `race` (X_7), `union` (X_8), `fulltime` (X_9), `tax.status` (X_{10}), `household.sum` (X_{11}), `n.employee` (X_{12}), `country.birth` (X_{13}), `citizenship` (X_{14}), `self.employed` (X_{15}), and `weeks.worked` (X_{16}).

hour while the average wage of women is only $8.03 per hour. Interestingly the IT tree has also identified a subgroup, Group IV, in which men are underpaid compared to women. This occurs in the following industries (X_5): 1 - Agriculture; 6 - Entertainment; 8 - Forestry and fisheries; 9 - Hospital services.

Finally, Figure 4(a) displays the computed variance importance scores for each of the 16 covariates. In this calculation, $B = 2,000$ bootstrap samples are used and the number of randomly selected variables, m_0, is set at 4 when splitting data. It can be seen that `industry` (X_5) and `household.sum` (X_{11}) stand out as the most important factors that contribute to the pay gap between males and females, followed by `occupation` (X_6), `tax.status` (X_{10}), `marital` (X_4), and then `age` X_1. This finding is consistent with the tree structure in Figure 3, except for that `tax.status` (X_{10}) and `marital`

	Male		Female		Two-Sample t Test	
	size	average	size	average	(computed from the test sample)	
Node	(n_1)	wage $(\bar{y}_1)$	(n_0)	wage $(\bar{y}_0)$	t	p-value
I	3220	12.29	1311	8.03	13.449	$< 0.0001^\star$
II	3800	10.11	5326	8.25	7.678	$< 0.0001^\star$
III	449	8.08	946	9.15	-0.306	0.7598
IV	501	9.43	1049	11.76	-5.560	$< 0.0001^\star$

Table 4: Summary statistics for the four final subgroups of the CPS data. Note that $(n_1, \bar{y}_1, n_0, \bar{y}_0)$ are computed using the whole data. The two-sample t test statistics are calculated from the test sample $\mathcal{L}_3$. The associated p-values are obtained for two-sided tests.

(X_4) have been masked out. The simple feature selection technique is also used to determine the importance of each covariate. Figure 4(b) depicts the resulting logworths. While both plots show some similarities, one should keep in mind that evaluation based on the latter approach is rather restrictive.

5. Discussion

In data analysis, it is important to distinguish two types of interactions. If there is no directional change in terms of the comparison, that is, $(\mu_1^L - \mu_0^L) \cdot (\mu_1^R - \mu_0^R) > 0$, the interaction is said to be *quantitative*; otherwise, it is termed as *qualitative*. The presence of qualitative interactions causes more concerns than quantitative ones (see, e.g., Gail and Simon, 1985). The results from the IT procedure can help to address this issue. For instance, the qualitative interaction in the the CPS example is obviously present among the final four subgroups.

The proposed IT procedure is applicable to a number of different areas. For example, the IT structure can help identify the most and least effective subgroups for the investigational medicine. If the new medicine shows an overall effect that is significant, and, if even in the least effective subgroup under examination, it does not present any harmful side effects, then its release may be endorsed without reservation. In trials where the proposed compound is not found to be significantly effective, the tree-structured subgroup analysis may identify sub-populations that contribute to the failure of the compound. Information gained in this manner could provide the basis for establishing inclusion/exclusion criteria in future trials, and as such, could be of considerable value to the existing efforts in synthesizing compounds to fight deadly diseases such as cancer and HIV/AIDS. We believe that efforts along these lines make subgroup analysis an efficient and valuable tool for research in many application fields.

Acknowledgments

The authors would like to thank the editor and three anonymous referees, whose insightful comments and constructive suggestions have greatly improved a preliminary version of this manuscript. Hansheng Wang's research is supported in part by a NSFC grant (#10771006) and a grant from Microsoft Research Asia.

Appendix A. Categorical Splits

The following theorem provides motivation and justification for the computationally efficient strategy of "ordinalizing" categorical covariates in interaction trees as discussed in Section 2.1. It is analogous to Theorem 9.6 of CART (Breiman et al., 1984, pp. 274–278) yet makes a stronger statement.

Consider splits based on a categorical covariate X, whose possible values range over a finite set $C = \{c_1, \ldots, c_r\}$. Then any subset $A \subset C$, together with its complement $A' = C - A$, induces a partition of the data into $t_L = \{(y_i, \mathrm{trt}_i, \mathbf{x}_i) : x_{il} \in A\}$ and $t_R = \{(y_i, \mathrm{trt}_i, \mathbf{x}_i) : x_{il} \in A'\}$. In the setting of interaction trees, the splitting rule seeks an optimal partition $(A + A')$ to maximize the difference in treatment effect between two child nodes $\left\{(\mu_1^L - \mu_0^L) - (\mu_1^R - \mu_0^R)\right\}^2$, where $\mu_1^L = \mu_{A1} = E\left(y \,|\, x_{il} \in A, \mathrm{trt}_i = 1\right)$ denotes the treatment mean in the left child node, and a similar definition applies for the other μ's. For the sake of convenience, we assume $(\mu_1^L - \mu_0^L) > (\mu_1^R - \mu_0^R)$ so that an optimal split maximizes $(\mu_1^L - \mu_0^L) - (\mu_1^R - \mu_0^R)$.

Theorem 1. *Let $\mu_{ck} = E(y \,|\, X = c, trt = k)$ for any element $c \in C$ and $k = 0, 1$. If $(A + A')$ forms an optimal partition of C, then we have*

$$\mu_{c_1 1} - \mu_{c_1 0} \geq \mu_{c_2 1} - \mu_{c_2 0},$$

for any element $c_1 \in A$ and $c_2 \in A'$. ■

Proof. Define

$$d_1 = \min_{c \in A} (\mu_{c1} - \mu_{c0}) \quad \text{and} \quad d_2 = \max_{c \in A'} (\mu_{c1} - \mu_{c0}).$$

Accordingly, it suffices to show that $d_1 \geq d_2$.

To proceed, let

$$\begin{aligned}
A_1 &= \{c \in A : \ \mu_{c1} - \mu_{c0} = d_1\} \\
A_2 &= \left\{c \in A' : \ \mu_{c1} - \mu_{c0} = d_2\right\} \\
A_3 &= \{c \in A : \ \mu_{c1} - \mu_{c0} > d_1\} \\
A_4 &= \left\{c \in A' : \ \mu_{c1} - \mu_{c0} < d_2\right\}.
\end{aligned}$$

Thus $A = A_1 + A_3$ and $A' = A_2 + A_4$. We reserve a generic notation $d = \mu_1 - \mu_0$ for the treatment effect. Define

$$\begin{aligned}
d_i &= \sum_{c \in A_i} (\mu_{c1} - \mu_{c0}) \, \Pr\{X_j = c\}, \\
d_{ij} &= \sum_{c \in A_i \cup A_j} (\mu_{c1} - \mu_{c0}) \, \Pr\{X_j = c\}, \\
d_{ijk} &= \sum_{c \in A_i \cup A_j \cup A_k} (\mu_{c1} - \mu_{c0}) \, \Pr\{X_j = c\}.
\end{aligned}$$

for $i \neq j \neq k = 1, 2, 3, 4$. We also introduce notation $Q_i = \Pr\{X_j \in A_i\}$ for $i = 1, 2, 3, 4$. It can be verified that

$$d_{ij} = \frac{d_i Q_i + d_j Q_j}{Q_{ij}} \ \text{ with } \ Q_{ij} = Q_i + Q_j$$

and

$$d_{ijk} = \frac{d_i Q_i + d_j Q_j + d_k Q_k}{Q_{ijk}} \text{ with } Q_{ijk} = Q_i + Q_j + Q_k.$$

Since partition $(A + A')$ provides an optimal partition, $d_{13} - d_{24} > 0$ reaches the maximum among all possible partitions. In particular, for partition $(A_3 + A_3')$, we have $d_{13} - d_{24} \geq d_3 - d_{124}$, which can be shown equivalent to

$$d_1 \geq (1 - Q_3)\, d_{13} + Q_3\, d_{24}, \tag{4}$$

by first plugging in the following two relationships

$$\begin{aligned} d_3 &= \frac{Q_{13} d_{13} - Q_1 d_1}{Q_3} \\ d_{124} &= \frac{Q_{24} d_{24} + d_1 Q_1}{Q_{124}} \end{aligned}$$

and then simplifying. Similarly, for partition $(A_4' + A_4)$, we can establish

$$Q_4\, d_{13} + (1 - Q_4)\, d_{24} \geq d_2 \tag{5}$$

starting with $d_{13} - d_{24} \geq d_{123} - d_4$.

Now suppose that $d_2 \geq d_1$. It follows from (5) and (4) that

$$\begin{aligned} & Q_4\, d_{13} + (1 - Q_4)\, d_{24} \geq d_2 \geq d_1 \geq (1 - Q_3)\, d_{13} + Q_3\, d_{24} \\ \Longrightarrow\quad & \{Q_4\, d_{13} + (1 - Q_4)\, d_{24}\} - \{(1 - Q_3)\, d_{13} + Q_3\, d_{24}\} \;\geq\; d_2 - d_1 \\ \Longrightarrow\quad & Q_{12}(d_{24} - d_{13}) \;\geq\; d_2 - d_1, \text{ since } 1 - Q_3 - Q_4 = Q_1 + Q_2 = Q_{12}. \end{aligned}$$

However, $0 > d_{24} - d_{13}$ leads to $0 > d_2 - d_1$, which contradicts the condition $d_1 \leq d_2$. Thus we must have $d_1 > d_2$.

The theorem also holds in extreme cases such as $d_1 = d_3$ or $d_2 = d_4$, etc. The proofs are omitted.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Czaki, editors, *2nd Int. Symp. Inf. Theory*, pages 267–281. Budapest: Akad Kiado, 1973.

S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 255:1064–1069, 2000.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

S.-C. Chow and J.-P. Liu. *Design and Analysis of Clinical Trials: Concepts and Methodologies*. Hoboken, NJ: Wiley-Interscience, 2004.

A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition. *Computational Statistics and Data Analysis*, 4:185–204, 1986.

CPMP Working Party on Efficacy on Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorisations for medicinal products: Note for guidance. *Statistics in Medicine*, 14:1659–1682, 1995.

M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41:362–372, 1985.

I. Kononenko and S. J. Hong. Attribute Selection for Modeling. In: *Future Generation Computer Systems*, 13:181–195, 1997.

S. W. Lagakos. The challenge of subgroup analyses - reporting without distorting. *The New England Journal of Medicine*, 354:1667–1669, 2006.

M. Leblanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467, 1993

A. D. R. McQuarrie and C.-L. Tsai. *Regression and Time Series Model Selection*. Singapore: World Scientific, 1998.

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

A. Negassa, A. Ciampi, M. Abrahamowicz, S. Shapiro, and J.-F. Boivin. Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15:231–239, 2005.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

P. Sleight. Debate: Subgroup analyses in clinical trials: Fun to look at, but don't believe them! *Current Controlled Trials on Cardiovascular Medicine*, 1:25–27, 2000.

X. G. Su, T. Zhou, X. Yan, J. Fan, and S. Yang. Interaction trees with censored survival data. *The International Journal of Biostatistics*, vol. 4 : Iss. 1, Article 2, 2008. Available at: `http://www.bepress.com/ijb/vol4/iss1/2`.

R. Tibshirani and K. Knight. The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B*, 61:529–546, 1999.

L. Torgo. A study on end-cut preference in least squares regression trees. Proceedings of the 10th Portuguese Conference on Artificial Intelligence. *Lecture Notes In Computer Science*, 2258:104–115. Springer-Verlag: London, UK, 2001.

J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998.