EXTENSIONS OF REGRESSION TREES FOR SUBGROUP IDENTIFICATION

A Dissertation

by

DOOWON CHOI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Li Zeng
Committee Members,	Yu Ding
	Alaa Elwany
	Xia Hu
Head of Department,	Lewis Ntaimo

May 2021

Major Subject: Industrial Engineering

ABSTRACT

Effective analysis is a key to the science of data analytics. Substantial advancement in data analytics and science has been made. Yet, there is still a rationale and validity for further research and more studies because the existing popular subgroup identification models, such as regression trees, are not effective in some cases. This dissertation is a serious endeavor to tackle those cases and devise better subgroup identification models.

Regression tree models have been widely used for subgroup identification in various domains such as social sciences, education, and healthcare informatics. However, a direct application of regression trees cannot satisfy the specific needs and may miss actually existing subgroups or identify misleading subgroups, because of challenging situations in practice. This dissertation focuses on modifying and extending regression trees for subgroup identification to address some uncharted situations, including *i*) developing correlation trees for cases where correlation, instead of regression, is of interest, *ii*) developing robust logistic regression trees to address outlier problems, and *iii*) exploring the potentials of generalized extreme value regression trees and Firth's logistic regression trees for modeling imbalanced class data.

This research is an interdisciplinary study on the interaction of advanced statistical modelling and machine learning approaches to identify heterogeneous subgroups to conquer the challenges in various fields and practices. The proposed models provide tangible insights, theories, and exploratory tools for subgroup identification. The research is expected to be widely applicable to various fields such as personalized medicine and

optimal psychological interventions where subgroup analysis is the main concern. The potential impact of this research is intended for academia and industry and society in general.

DEDICATION

I dedicate this dissertation to my parents.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all people whose help has made it possible to accomplish my Ph.D. degree.

I am and always will be indebted to my academic advisor, Dr. Li Zeng. I consider myself very being blessed to have her as my advisor. She led me into the world of research and trained me as an independent researcher with thoughtful encouragement and careful supervision. When my progress was slow and I made mistakes, she kept her faith in me and was willing to share her ideas and directions for resolving research questions. Without her sacrificial commitment and sustained guidance, it would have been impossible to complete this dissertation.

I also sincerely thank my committee member, Dr. Yu Ding, for his constructive suggestions and enlightening comments on my research. Through joint lab meetings every semester, I have learned a great deal from his insightful views of research. I would also like to thank Dr. Alaa Elwany for his precious comments on my research and guidance for the role and responsibility of an teaching assistant. I also would like to express gratitude to Dr. Xia Hu for his useful comments and inspiring suggestions on the directions of my future study.

I must also thank my friends in the Lab and fellow graduate students: Qian Wu, Hoon Hwangbo, Yanjun Qian, Ahmed Aziz Ezzat, Imitiaz Ahmed, Shilan Jin, Abhinav Prakash, Adaiyibo Kio, Jiaxi Xu, and Seyed Mohammad Hossein. They carefully listened to me when I needed listening ears, and gave me their hands when I needed them. They have made my days in College Station a warm and happy memory which I will never forget.

More than anything, I have great gratitude and love for my parents and my brother. They have unconditionally supported me with all their means and wishes for the completion of my study. I am forever thankful and indebted to their love and support.

Lastly, I would like to say many thanks to Pastor and friends at Korean Church of A&M for their prayers and friendships.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Li Zeng, Professor Yu Ding, and Professor Alaa Elwany of the Department of Industrial and Systems Engineering and Professor Xia Hu of the Department of Computer Science and Engineering.

The data analyzed for Chapter III was provided by Susan Seidensticker, Dr. Carlos Clark, Dr. Lindsey Sonstein, Rick Trevino, and Dr. Gulshan Sharma in the University of Texas Medical Branch, Galveston. TX.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a teaching assistantship from Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

I.1 Motivation

Proper data collection and its pursuant effective analysis are keys to the science of data analytics. Substantial advancements in data analytics and science have been made. One of the advancements is to handle heterogeneity of observations in data. For example, in healthcare data, there is substantial heterogeneity of patients, together with a large number of variables and complex data structures. A global model is often not adequate to explain such complex data. This necessitates a revision of the comprehensive model so that the new model can be flexible enough to interpret and accommodate the heterogeneity to help policymakers and clinicians in making decisions tailored to the subpopulations. Many studies have been done for subgroup identification, but this topic is still an underdeveloped area and more research is needed to tackle special problems in reality. This dissertation is a serious endeavor to this area.

The general topic of subgroup identification has attracted much attention in the clinical trial and biostatistics community [1]. Basically, subgroup identification aims to identify the right patients for a particular treatment and thus discover subpopulations that would have enhanced benefits from the treatment. Likewise, subgroup identification is able to find the right treatment for the particular patient and thus identify optimal treatment policy or plan for a given subpopulation. This implies that heterogeneous treatment effect exists and the effect can be characterized by the interaction of the treatment with patient's characteristics. For those purposes, a number of methods for subgroup identification have been developed in the field of personalized medicine and clinical drug development [2-12].

However, the idea of subgroup identification is not limited to clinical trial applications. In manufacturing systems, we can identify the heterogeneous effects of the particular machine/process (i.e., the counterpart of treatment in clinical trial applications) on defective products with interactions of other machines/processes. Also, practitioners can discover sets of processes that benefit from a particular maintenance plan most, which can lead to substantial increase in yields in the manufacturing line. In healthcare applications, hospitals can single out groups of patients with different morbidities, mortalities or readmission rates and further achieve optimal resource allocation and best practices by group-specific care management. Thus, this dissertation does not confine subgroup identification to the topic of clinical trial applications, but instead aims to devise better subgroup identification models applicable to many other applications.

This dissertation benchmarks regression trees for subgroup identification. As a class of decision trees, regression trees are able to handle nonlinear relationships and heterogeneity of data by recursively partitioning the covariate space into subgroups and fitting a regression model for each resulting subgroup. By identifying subgroup-specific regression models, a regression tree not only explains the data better, but also preserves the good interpretability of regression models. An example of regression tree is given in Figure I.1, where there are two split variables (gender and age), $\boldsymbol{X}$ is the predictors (e.g., patient's risk factors), y is the response (e.g., outcome of medical procedure) and $f(\boldsymbol{X}; \boldsymbol{\theta})$

is a regression model with coefficients $\boldsymbol{\theta}$. Note that some regression trees use every covariate as both predictor and split variable, while others specify the role of covariates (only for splitting, only for regression, or for both splitting and regression). In this example, the split variables and predictors are exclusive with each other. In Figure I.1, the population is divided into three subgroups according to gender and age, producing three regression models with different coefficients. These subgroup-wise models reflect the heterogeneous effect of patient's risk factors on outcomes of the medical procedure.

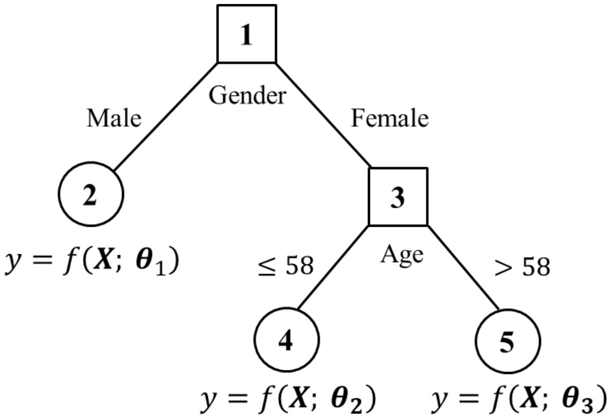

1
Gender
Male
Female
2
$y = f(\boldsymbol{X};\ \boldsymbol{\theta}_1)$
3
Age
≤ 58
> 58
4
5
$y = f(\boldsymbol{X};\ \boldsymbol{\theta_2})$
$y = f(\boldsymbol{X};\ \boldsymbol{\theta_3})$

Figure I.1 Illustrative example of regression tree

Linear regression trees and logistic regression trees are two examples of regression trees. As mentioned above, these tree models have been widely used for subgroup identification in domains like social sciences, biomedical engineering and healthcare studies. However, existing regression trees are not able to cope with some special aspects of practical problems in those domains. In this dissertation, we consider three of those special issues and extend the existing tree models to solve them: 1) subgrouping based on correlation instead of regression, 2) addressing outlier problems, and 3) modelling imbalanced class data.

I.2 Research objective and challenges

I.2.1 Correlation trees for subgroup identification in brain-behavior analysis

Correlation is the most widely used measure for quantifying strength of the relationship between two variables. It is natural that correlation depends on the condition of other covariates. For example, in neural correlates study aiming to understand the neural basis of human experiences in cognitive processes, the brain-behavior correlation is a primary measure for study. As this correlation depends on subject-related covariates such as age and gender, it is common that simple subgroups (e.g., old vs. young and male vs. female) are specified according to common sense or prior knowledge from the literature. However, such a primitive approach of subgroup identification has many problems. First, it suffers from difficulties in forming subgroups by continuous covariates like age. The appropriate cutting point to different age groups (e.g., young vs. old) is usually not obvious and depends on the nature of the study. Second, a single covariate, either age or gender, is used in most studies, which is limited for explaining the correlation. Other related covariates (e.g., education, health conditions) should be considered for better subgrouping. Finally, interaction of covariates may also affect the brain-behaviour correlation and thus need to be considered. This study proposes an approach called correlation tree for automatic subgroup identification in such correlation analysis and provide meaningful objective functions to meet various needs in practice.

I.2.2 Addressing outliers in subgroup learning for outcome data in healthcare

Measuring healthcare outcomes has become highly essential for quality of care assessment, improvement and evidence-based practice. Outcome measures usually have different relationships with covariates such as physiological and treatment variables among patients. Therefore, subgroup-wise modelling is indispensable for healthcare professionals. Logistic regression trees can serve this purpose by recursively partitioning the covariates space into a number of subgroups in such a way that a single logistic regression adequately fits the data in each subgroup. However, real-world data do not always conform to the familiar/normal structures and can be contaminated by aberrant observations such as outliers.

Outliers often exist in real data. In general, outliers are defined as observations that are extremely deviated from the bulk of data (Hawkins, 1980). They may have substantial effect in regression analysis, especially when the least squares method is used for model fitting. The least squares method aims to minimize the sum of residuals, so it tends to avoid large residuals by nature. Consequently, outliers are accommodated at the expense of poor fit for the majority of data. To the best of our knowledge, the outlier problem has not been studied in the context of tree models. As mentioned before, tree models involve two aspects: covariate space splitting (i.e., subgroup identification) and model fitting for each subspace. The effect of outliers on the fitting aspect is similar to that in regression (i.e., causing a poor fit), but the effect of outliers on the splitting aspect is still unknown. It is believed that classification trees are relatively robust against outliers as their splitting criteria are functions of proportions of classes which are not highly sensitive to outliers. However, regression trees where the splitting is based on variance of data may be affected seriously. For example, they may fail to split a node that should be split due to outliers, resulting in misleading subgroups. This dissertation considers the outlier problem for subgroup identification in logistic regression trees and develops a robust logistic regression tree approach against the outlier effect.

I.2.3 Subgroup identification for imbalanced class data

With years of quality improvement efforts in many applications, the number of adverse outcomes like mortality is gradually decreasing. It has been known that conventional binary outcome modelling methods are meant to favor the majority class and tend to underestimate the probability of the minority class in prediction. Thus, most methods have primarily prioritized the improvement on prediction accuracy of the minority class. However, the methods designed for improving prediction accuracy may not work well in the context of subgroup identification. This dissertation tries to better understand such an imbalanced class issue in the context of subgroup identification beyond prediction and proposes new binary regression trees for subgroup identification for imbalanced class data.

I.3 Organization of the dissertation

This dissertation is organized as follows. Chapter II develops a correlation tree for subgroup identification. The correlation tree automatically identifies subgroups with different correlations through systematic unbiased split variable selection and the estimation of the optimal cutpoint for the selected split variable. In particular, the proposed correlation trees handle both linear and non-linear correlation measures and provide three types of practical objective functions to meet needs in various applications.

Chapter III is dedicated to studies on the outlier problem in the context of subgrouping by logistic regression trees. This study is to reveal the effect of outliers on subgroup learning by simulation and propose a logistic regression tree robust to outliers. To demonstrate the effectiveness of the proposed robust logistic regression tree, we incorporate down-weighting and outlier detection method with logistic regression trees and compare the performance of the three methods.

Chapter IV concerns the imbalanced class problem in subgroup identification. Two binary regression trees are proposed and their performance is compared with that of the conventional logistic regression tree under different degrees of balance by simulation. Through this study, we provide better understanding on the three methods and their advantages in modeling imbalanced class data.

Chapter V summarizes this dissertation and highlights its contributions. We also discuss potential future directions of this dissertation study.

CHAPTER II

A RECURSIVE PARTITIONING APPROACH FOR SUBGROUP IDENTIFICATION IN BRAIN-BEHAVIOUR CORRELATION ANALYSIS[*]

In neural correlates studies, the goal is to understand the brain-behaviour relationship characterized by correlation between brain activation responses and human behaviour measures. Such correlation depends on subject-related covariates such as age and gender, so it is necessary to identify subgroups within the population that have different brain-behaviour correlations. The subgrouping is made by manual specification in current practice, which is inefficient and may ignore potential covariates whose effects are unknown in the literature. This study proposes a recursive partitioning approach, called correlation tree, for automatic subgroup identification in brain-behaviour correlation analysis. In constructing a correlation tree, the split variable at each node is selected through an unbiased variable selection method based on partial correlation test, and then the optimal cutpoint of the selected split variable is determined through exhaustive search under an objective function. Three types of meaningful objective functions are considered to meet various practical needs. Results of simulation and application to real data from optical brain imaging demonstrate effectiveness of the proposed approach.

[*] Reprinted with permission from "A recursive partitioning approach for subgroup identification in brain-behaviour correlation analysis" by Doowon Choi, Lin Li, Hanli Liu and Li Zeng, Pattern Analysis and Applications, 23(1), 161-177, Copyright © 2019, Springer-Verlag London Ltd., part of Springer Nature.

II.1 Introduction

Neural correlates is widely studied in disciplines like neuroscience, biomedical engineering, and brain research [13-15]. The brain is a complicated system comprised of an infinite number of neurons responsible for various tasks. The aspiration of neural correlates studies is to understand the neural basis of human experiences in cognitive processes such as decision making by revealing the brain-behaviour relationship. Typically, changes in brain parameters (e.g., blood flow, electrical signal) during a process are acquired by imaging techniques such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), and then correlation between those data and human behaviour measures is studied. As the brain-behaviour correlation depends on subject-related covariates such as age and gender, it is common in such studies that simple subgroups (e.g., old vs. young, male vs. female) are specified according to common sense or prior knowledge from literature, and correlation analysis is conducted for each subgroup separately [16-18]. This will reveal population heterogeneity in cognitive patterns as well as better explain the brain-behaviour correlation.

This primitive approach of subgroup specification has many problems. First, it suffers from difficulties in forming subgroups by continuous covariates like age. The appropriate cutting point to different age groups (e.g., young vs. old) is usually not obvious and depends on the nature of the study; specification based on convention in the literature may not work. Second, a single covariate, either age or gender, is used in most studies, which is limited for explaining the correlation. Other related covariates (e.g., education, occupation, health conditions) should be considered for better subgrouping. In presence of more than one covariates, an automatic approach is needed to select important covariates and group the population based on the selected covariates. Finally, interaction of covariates may also affect the brain-behaviour correlation and thus need to be considered. For example, it is often the case that certain difference between males and females depends on age, so incorporating the interaction of gender and age will make results more interpretable. In summary, there needs a general, advanced data analytics framework for automatic subgroup identification in the brain-behaviour correlation analysis.

Recursive partitioning is a popular approach for subspace segmentation in regression and classification analysis, known as decision trees. Here we adapt it to the considered subgroup identification problem. The proposed approach, called correlation tree, identifies subgroups with different brain-behaviour correlations via recursive binary partitioning. In addition to the identified subgroups, this approach also provides a convenient and rigorous way to find important covariates, as split variables in the tree, associated with the correlation through an unbiased variable selection method. Moreover, we consider both linear and nonlinear correlations and three types of meaningful objective functions in tree splitting to meet various needs in practice. It is worth mentioning that although the proposed approach is illustrated using the neural correlates studies in this paper, it can be easily applied to other correlation analysis.

The remainder of this paper is organized as follows. Section II.2 reviews related literature, including basics of correlation measures and concept and popular algorithms of decision trees. Section II.3 describes the proposed correlation tree. A simulation study is presented in Section II.4 to validate the variable selection method as a critical component of the proposed approach. Section II.5 applies the approach to real data from an optical brain imaging study. Section II.6 concludes the chapter and discusses future research.

II.2 Literature review

II.2.1 Correlation measures

Correlation is the most widely used measure for quantifying strength of the relationship between two variables. It is measured by a statistic called correlation coefficient which is a dimensionless quantity in the range [–1, +1]. A zero correlation coefficient indicates that no relationship exists between the two variables, while a value being –1 or +1 indicates a perfect (negative or positive) relationship. If the correlation coefficient is positive, it means that when one variable increases or decreases, the other one also increases or decreases, i.e., they follow the same trend. A negative correlation coefficient means that they follow opposite trends. In neural correlate studies, the sign of correlation between brain parameter and behavior measure has a specific interpretation regarding human cognitive patterns, as shown in Section II.5.

There are two types of correlation measures depending on the nature of the relationship: linear and nonlinear correlation measures. The first one measures how strongly two variables are linearly proportional to each other. A popular measure of this type is Pearson’s correlation coefficient. The second one concerns non-linear or general correlation. Examples of this type are Spearman’s rank correlation coefficient and Kendall’s Tau correlation coefficient [19]. They assess correlation using rank values of variables and measure how well the relationship can be described by a monotonic function. The Spearman's takes a similar form as Pearson's correlation coefficient (i.e., covariance of the rank variables divided by their standard deviations), while the Kendall's considers the concordance of the rank variables in every pair of observations. The more the concordant pairs are, the higher the correlation is.

As an example, Figure II.1 illustrates the use of different correlation measures on simulated data of variables X_1, X_2. In Figure II.1(a) where the data exhibit a linear relationship, Pearson's correlation coefficient (value = 0.811) and Spearman's rank correlation coefficient (value = 0.792) are similar. In Figure II.1(b) where the data have a monotonic rather than linear trend, the Spearman's (value = 0.842) is greater than the Pearson's (value = 0.784). It is also robust to outliers as it is based on rank values.

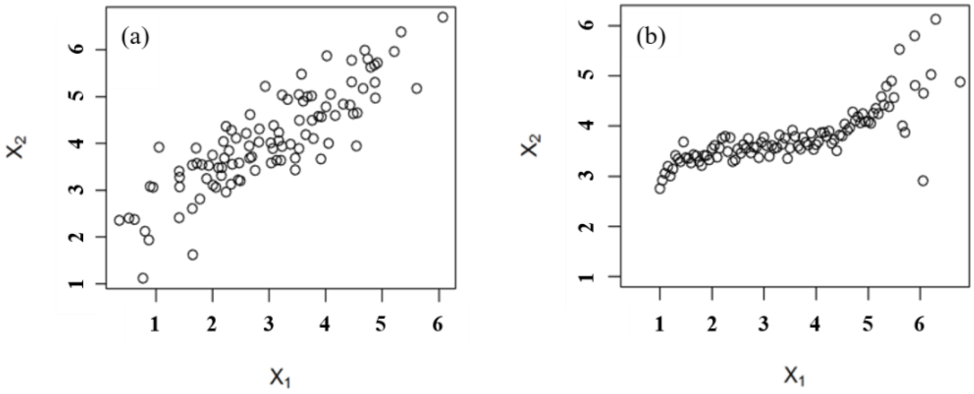


Figure II.1 Illustration of correlation measures

II.2.2 Decision trees

An example of decision tree is given in Figure II.2, where there are two predictors (age and blood pressure) and the response is outcome of surgery (D=died, A=alive). As shown in the left panel of Figure II.2, the decision tree splits the predictor space by age and blood pressure such that the resulting subregions contain the most homogeneous surgical outcomes, and thus prediction can be performed within each subregion separately. The graphical representation of the tree model is shown in the right panel. The terminal nodes or leaves (i.e., nodes 2, 4, 5) represent the resulting subregions; the root node (i.e., node 1) and the internal node (i.e., node 3) indicate how the predictor space is split; and the segments of the tree that connect the nodes are called branches. In this example, the tree first splits the predictor space by blood pressure (less than or equal to 152 vs. greater than 152), and then further splits the region with blood pressure greater than 152 by age (younger than or equal to 58 vs. greater than 58), resulting in three subregions.

The tree model is constructed through a recursive partitioning procedure that divides each node into two (or more) subnodes according to a cutpoint of certain predictor. In Figure II.2, the cutpoint of blood pressure is 152, while that of age is 58. Optimal splitting, including the optimal split variable and the optimal cutpoint of the variable, is involved in each iteration. It aims to improve an objective function that measures the fitting of data. The splitting stops when a pre-specified stopping rule is achieved.

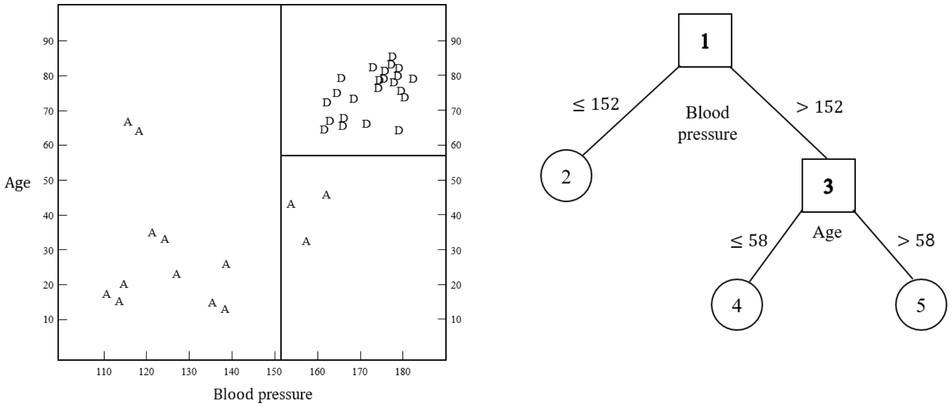


Figure II.2 Example of decision tree: recursively partitioned predictor space (left) and the graphical representation (right)

Depending on the type of the response variable, decision trees are categorized into two groups: classification trees for categorical response, and regression trees for continuous response. As the two variables involved in correlation analysis are continuous, regression trees are relevant to this study. There are two types of regression trees, piecewise-constant tree that uses a constant (i.e., mean of response values) for prediction in each subregion, and piecewise-linear regression tree that fits a linear regression model for the prediction. Popular algorithms of each type will be reviewed in detail in the following.

The ancestor of piecewise-constant trees is the Automatic Interaction Detection (AID) algorithm [20]. By defining impurity as the sum of squared prediction errors at a node, AID iteratively searches the split variable that minimizes the sum of impurities at the two subnodes. The algorithm is terminated when the reduction in impurity is below a pre-determined fraction of the initial impurity. Classification And Regression Tree (CART) [21] is one of the most popular piecewise-constant trees. The tree learning approach is the same as in AID, except for a different way to control tree size. Rather than relying on stopping rules that prevent a tree from growing in advance, CART first grows a tree as large as possible, and then cuts it back to find a sub-tree that has the lowest cross-validation error. Note that controlling tree size through stopping criteria is called pre-pruning, while the approach CART uses is called post-pruning. Both are meant to improve the prediction performance and interpretability of the tree model.

For piecewise-linear regression trees, Smoothed and Unsmoothed Piecewise POlynomial Regression Trees (SUPPORT) [22] is a typical algorithm. Rather than simultaneously selecting split variable and its cutpoint as CART does, SUPPORT first goes through a variable selection step and then finds the optimal cutpoint of the selected variable. The advantage of such a two-step approach is that it can avoid selection bias for a split, i.e., variables with more possible cutpoints have higher chance to be selected, an intrinsic problem of CART and similar algorithms [23-26]. For the variable selection, SUPPORT first computes the residuals of a linear regression model at the node using all possible split variables and then divides data into two groups according to signs of residuals. Next, for each split variable, this algorithm compares its mean and variance in the two groups by two-sample t-test and Levene test [27]. The idea is that these parameters, which characterize the variable's distribution, should not differ in the two groups if the fitted model is satisfactory. Finally, the variable that has most significant differences is chosen to split the node, and the cutpoint is the average of the two sample means.

An extended version of SUPPORT is the Generalized, Unbiased, Interaction Detection and Estimation (GUIDE) algorithm [28] which can be used for both classification and regression. GUIDE can handle categorical predictors and interaction of predictors which are main limitations of SUPPORT. Unlike other regression trees, GUIDE first specifies the role of each predictor variable (only for splitting, only for regression, or for both splitting and regression) before starting tree building. For unbiased variable selection, GUIDE conducts chi-square independent test which allows split variables to be selected with equal probability. Specifically, at each node, a linear regression model is fitted using all predictor variables pre-defined only for regression or for both splitting and regression, and residuals are computed. Then, for each predictor variable that serves only for splitting or for both splitting and regression, chi-square test is conducted to identify its association with the signs of residuals, which is essentially a lack of fit test. As chi-square test is designed for categorical variables, discretization is needed for continuous splitting variables in this step. The variable with the smallest p-value in the tests will be selected as the split variable, and the optimal cutpoint of this variable will be found by minimizing the total sum of squared residuals in the two subregions.

II.3 The proposed approach

Let X_1 and X_2 be two continuous variables, the correlation ρ of which is of interest, and $\mathbf{Z} = \{Z_1, Z_2, \dots, Z_k\}$ be the set of covariates. For example, in neural correlates studies, X_1 is a brain parameter (e.g., change in brain blood flow), X_2 is a behaviour measure (e.g., response time to a stimulus), and $\mathbf{Z}$ consists of gender, age, education, etc. This study will build a correlation tree model using the covariates as split variables to identify subgroups with different correlations of X_1 and X_2.

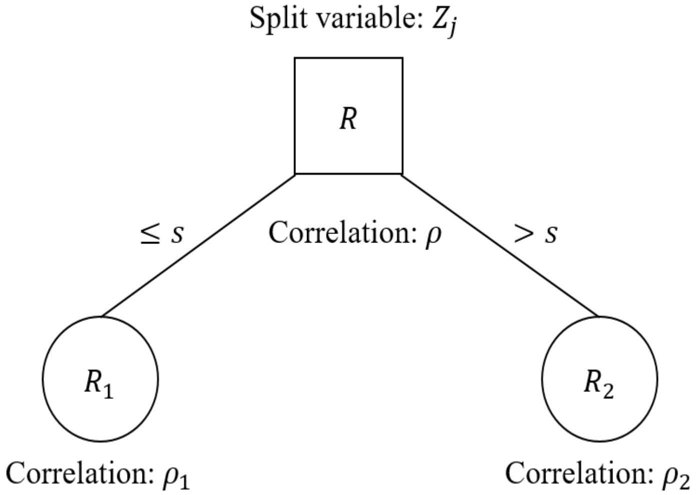


Figure II.3 Partitioning at each node in the proposed correlation tree

Like a regression tree, the correlation tree is constructed iteratively by splitting a node into several subnodes in each iteration. A binary splitting (i.e., two subnodes) is commonly used in regression trees literature, which will be followed in this study. The space partitioning at each node in the correlation tree is illustrated in Figure II.3. Formally, the problem is defined as follows: Given a split variable Z_j, $1 \leq j \leq k$, and a cutpoint s of this variable, the current node (region) R is partitioned into two subregions

$$R_1 = \{\mathbf{Z} | Z_j \leq s\}, \quad R_2 = \{\mathbf{Z} | Z_j > s\}.$$

Assume n samples $\{(x_{11}, x_{21}, \mathbf{z}_1), (x_{12}, x_{22}, \mathbf{z}_2), \dots, (x_{1n}, x_{2n}, \mathbf{z}_n)\}$ are available at the current node. Then data of X_1 and X_2 falling into these two regions are

$$\{(x_{1i}, x_{2i})\}_{i:\mathbf{z}_i \in R_1}, \quad \{(x_{1i}, x_{2i})\}_{i:\mathbf{z}_i \in R_2}.$$

Among all possible split variables and all possible cutpoints of each variable, we want to find the split variable and its cutpoint such that an objective function

$$\Psi = f(\rho_1, \rho_2) \tag{II.1}$$

will be optimized, where ρ_1 and ρ_2 are correlations of X_1 and X_2 in the two subregions, and f is a function of these two correlations which has meaningful interpretation (e.g., average of them). Examples of the objective function will be given in Section II.3.2.

We will consider Pearson's correlation coefficient and Spearman's rank correlation coefficient as correlation measure. Pearson's correlation coefficient of X_1 and X_2 is defined as

$$\rho = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1) \cdot Var(X_2)}}, \qquad \hat{\rho} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}}.$$

Here $Cov(X_1, X_2)$ is covariance of X_1 and X_2, and $Var(X_1)$ and $Var(X_2)$ are variances of X_1 and X_2. $\hat{\rho}$ is the estimate of the correlation at the current node using available data, where $\bar{x}_1$ and $\bar{x}_2$ are sample means of the two variables. Spearman's rank correlation coefficient is simply the Pearson's using rank data [29]

$$\rho = \frac{Cov\left(r_{X_1}, r_{X_2}\right)}{\sqrt{Var\left(r_{X_1}\right) \cdot Var\left(r_{X_2}\right)}}, \qquad \hat{\rho} = \frac{\sum_{i=1}^{n}\left(r_{x_{1i}} - \bar{r}_{x_{1i}}\right)\left(r_{x_{2i}} - \bar{r}_{x_{2i}}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_{x_{1i}} - \bar{r}_{x_{1i}}\right)^2 \sum_{i=1}^{n}\left(r_{x_{2i}} - \bar{r}_{x_{2i}}\right)^2}},$$

where $r_{X_1}$and r_{X_2} are the rank variables of X_1 and X_2. The tree is a linear correlation tree (LCT) when the Pearson's is used and a non-linear correlation tree (NCT) when the Spearman's is used.

Our proposed algorithm solves the partitioning problem illustrated in Figure II.3 in two steps: first selecting the optimal split variable Z_j^* and then finding the optimal cutpoint s^* of the selected variable. As in SUPPORT and GUIDE, the variable selection step is meant to eliminate selection bias on split variables. When a tree searches the optimal split variable and optimal cutpoint simultaneously, covariates with more possible cutpoints are favoured because they can generate bigger, finer solution spaces in optimization. For example, assume Z_1 is a continuous variable with L distinct values and Z_2 is a categorical variable with M categories. Then Z_1 has $(L-1)$ possible cutpoints, while Z_2 has $(2^{M-1}-1)$ possible cutpoints. If $(L-1) > (2^{M-1}-1)$, Z_1 is more likely to be selected than Z_2; otherwise Z_2 is more likely to be selected. To enable unbiased selection of split variables, a method based on partial correlation test is proposed to find the optimal split variable. The unbiasedness of this method is validated by simulations in

Section II.4. The optimal cutpoint of the selected variable will be obtained via exhaustive search to optimize a pre-defined objective function. Details of these two steps will be given as follows, together with how to control tree size to ensure good interpretability and an analysis of time complexity.

II.3.1 Variable selection

Intuitively, the optimal split variable Z_j^* at each node should be the covariate in the set $\{Z_1, Z_2, \ldots, Z_k\}$ that explains the correlation of X_1 and X_2 most. The proposed method for variable selection uses partial correlation test to find this variable. Before elaborating this method, we first introduce the concept of partial correlation. For each covariate Z_j, the partial correlation $\rho_{X_1X_2 \cdot Z_j}$ measures strength of the relationship between X_1 and X_2 after adjusting for the effect of Z_j [30]. More precisely, $\rho_{X_1X_2 \cdot Z_j}$ is the correlation of the remaining parts of X_1 and X_2 after partialing out the effect of Z_j on them. One can regard that the partial correlation is the reciprocal information between X_1 and X_2 that is not explained by Z_j. Figure II.4 illustrates this idea, where the partial correlation $\rho_{X_1X_2 \cdot Z_j}$ is represented by the blue area.

Given the above definition of partial correlation, we can draw the following insights: If $\rho_{X_1X_2 \cdot Z_j}$ is equal to the regular correlation $\rho_{X_1X_2}$, it means that Z_j is independent of the correlation of X_1 and X_2. The higher $\rho_{X_1X_2 \cdot Z_j}$ is, the less the covariate Z_j explains the correlation of X_1 and X_2, and vice versa. In other words, the magnitude of the partial correlation is inversely proportional to the extent to which Z_j explains the correlation of X_1 and X_2 which is represented by the red area in

Figure II.4. Therefore, the optimal split variable Z_j^* is supposed to have the smallest degree of partial correlation (i.e., the smallest blue area, or equivalently, the largest red area).

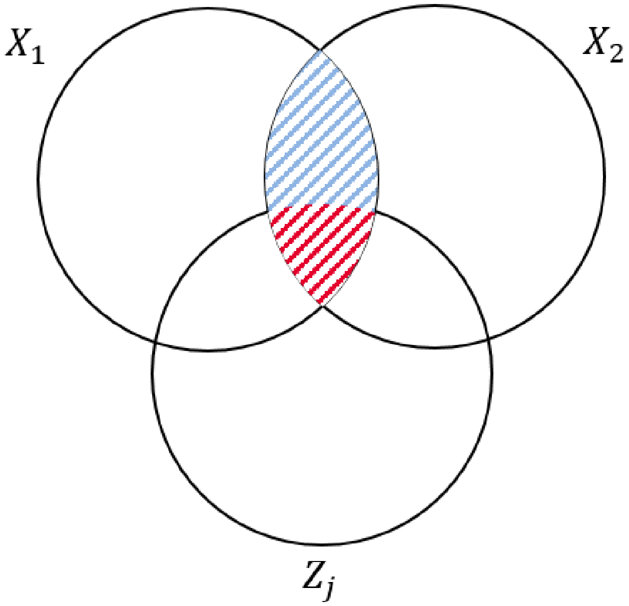


Figure II.4 Illustration of partial correlation

Based on this idea, the optimal split variable can be found by the following steps: First, calculate the sample partial correlation coefficient $\hat{\rho}_{X_1X_2\cdot Z_j}$ for each covariate Z_j using available data at the current node. Second, test whether the partial correlation is zero (i.e., H_0: $\rho_{X_1X_2\cdot Z_j} = 0\ vs. H_1$: $\rho_{X_1X_2\cdot Z_j} \neq 0$) using $\hat{\rho}_{X_1X_2\cdot Z_j}$. In general, the null distribution of $\hat{\rho}_{X_1X_2\cdot Z_j}$is complicated, so there is no simple test directly based on it. To solve this problem, we can apply the Fisher's $\mathcal{Z}$-transformation to the sample partial correlation

$$\mathcal{Z} = \frac{1}{2} ln\left(\frac{1+\hat{\rho}_{X_1X_2\cdot Z_j}}{1-\hat{\rho}_{X_1X_2\cdot Z_j}}\right).$$

It is known that

$$W = \sqrt{n - d - 3} \cdot \mathcal{Z} \tag{II.2}$$

approximately follows a standard normal distribution, where n is the number of samples and d is the number of adjusting variables [31]. H_0 is rejected if $|W| > \Phi^{-1}(1 - \alpha/2)$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution and α is the significance level. P-value of the test will be obtained. Finally, compare *p*-values of all covariates in the partial correlation test. The covariate with the largest *p*-value (indicating the least significant partial correlation) will be selected as the single optimal split variable. The algorithm is summarized in Table II.1.

Table II.1 The proposed algorithm for split variable selection

Algorithm 1

1. *Main effect test: For each covariate Z_j, fit a linear regression model of X_1 using it and compute the residuals $\epsilon_{X_1 \sim Z_j}$; similarly, obtain the residuals of X_2, $\epsilon_{X_2 \sim Z_j}$.*
2. *Calculate the sample partial correlation coefficient $\hat{\rho}_{X_1 X_2 \cdot Z_j}$ as the correlation of the two residuals.*
3. *Compute the test statistic W in Equation (II.1) and its corresponding p-value from a standard normal distribution.*
4. *Interaction effect test: For a pair of two covariates (Z_j, Z_l), fit a linear regression model of X_1 using their interaction and compute the residuals $\epsilon_{X_1 \sim Z_j Z_l}$; similarly, obtain the residuals of X_2, $\epsilon_{X_2 \sim Z_j Z_l}$.*
5. *Calculate the sample partial correlation coefficient $\hat{\rho}_{X_1 X_2 \cdot Z_i Z_j}$ as the correlation of the two residuals.*
6. *Compute the test statistic W in Equation (II.1) and its corresponding p-value from a standard normal distribution.*
7. *Compare the p-values of all covariates and interactions, and select the covariate with the largest p-value as the optimal split variable.*

Some related problems in implementing the algorithm are discussed as follows:

- *Calculation of partial correlation:* The sample partial correlation coefficient can be obtained by [32]

$$\hat{\rho}_{X_1X_2\cdot Z_j} = \frac{\hat{\rho}_{X_1X_2} - \hat{\rho}_{X_1Z_j}\hat{\rho}_{X_2Z_j}}{\sqrt{1-\hat{\rho}^2_{X_1Z_j}}\cdot\sqrt{1-\hat{\rho}^2_{X_2Z_j}}}, \tag{II.3}$$

  where $\hat{\rho}_{X_1Z_j}$ is the sample correlation of X_1 and Z_j and $\hat{\rho}_{X_2Z_j}$ is that of X_2 and Z_j. Algorithm 1 uses another way to find this quantity. It calculates the correlation of two residuals, $\epsilon_{X_1\sim Z_j}$, from regressing X_1 on Z_j, and $\epsilon_{X_2\sim Z_j}$, from regressing X_2 on Z_j [33]. Then, Equation (II.3) can be rewritten as

$$\hat{\rho}_{X_1X_2\cdot Z_j} = \frac{\sum_{i=1}^{n}\left[\left\{\left(x_{1i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_1\right)-\frac{1}{n}\sum_{i=1}^{n}\left(x_{1i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_1\right)\right\}\left\{\left(x_{2i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_2\right)-\frac{1}{n}\sum_{i=1}^{n}\left(x_{2i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_2\right)\right\}\right]}{\sqrt{\sum_{i=1}^{n}\left\{\left(x_{1i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_1\right)-\frac{1}{n}\sum_{i=1}^{n}\left(x_{1i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_1\right)\right\}^2\sum_{i=1}^{n}\left\{\left(x_{2i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_2\right)-\frac{1}{n}\sum_{i=1}^{n}\left(x_{2i}-\boldsymbol{g}_{ji}^T\hat{\boldsymbol{\beta}}_2\right)\right\}^2}} \tag{II.4}$$

  where x_{1i}, x_{2i} and z_{ji} are the ith observation of X_1, X_2 and Z_j, $\boldsymbol{g_{ji}} = \left[1, z_{ji}\right]^T$, and $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are the least squares estimates of parameters in the linear regression model of X_1 and X_2. Such a residual-based approach is more flexible than the formula in Equation (II.3) as it works for both continuous and categorical covariates. Note that when the covariate Z_j is a categorical variable, its correlation with X_1 or X_2 is not well defined, and thus Equation (II.3) cannot be used. In contrast, the residual-based approach relies on linear regression on Z_j which holds whether Z_j is continuous or categorical.

- *Partial correlation of categorical covariates:* When Z_j is a categorical covariate, dummy coding should be applied in linear regression [34]. Suppose Z_j has three categories A, B, C. Then, it can be replaced with two dummy variables (indicators), $\mathrm{D_A}$ and $\mathrm{D_B}$, each taking two possible values (0 and 1). If the observation of Z_j is A, then $\mathrm{D_A}$ is equal to 1, otherwise 0. The same applies for $\mathrm{D_B}$. Naturally both $\mathrm{D_A}$ and $\mathrm{D_B}$ equal to 0 indicates the third category C. Linear regression of X_1 or X_2 will be conducted using the two dummy variables as regressors. Note that for a categorical covariate with M categories, $M-1$ dummy variables will be used and thus *d* in Equation (II.2) takes a value of $M-1$.

- *Interaction effect test:* To enhance the performance of variable selection, interaction of covariates can be considered. This poses three changes on Algorithm 1: First, to find the sample partial correlation adjusted by the interaction of two covariates Z_j and Z_l using the residual-based approach, X_1 and X_2 should be regressed on the interaction $Z_j Z_l$. Second, *d* in Equation (II.2) takes a value of 1, $M-1$ and $(M_1-1)(M_2-1)$ for interaction between two continuous variables, a continuous variable and a categorical variable with M categories, and two categorical variables with M_1 and M_2 categories, respectively. Third, if an interaction has the largest *p*-value in the partial correlation test, the covariate in the interaction that has the larger *p*-value in the test for a single covariate will be selected.
- *Nonlinear correlation:* Equations (II.3) and (II.4) apply for both the Pearson's and Spearman's correlations. To obtain the Spearman's partial correlation, we simply transform the data into ranks and calculate the Pearson's partial correlation using the rank data. Note that when there are multiple observations of the same value (i.e., ties),

they are assigned the average of their original orders in the rank transformation. For example, for observations {1, 2, 3, 3, 5}, the corresponding ranks are {5, 4, 2.5, 2.5, 1}.

II.3.2 Finding optimal cutpoint

Once the optimal split variable Z_j^* is obtained, we will search the cutpoint s^* of this variable to optimize the objective function Ψ in Equation (II.1). In general, the objective function depends on interest in the specific research context. To meet needs in different contexts, three types of meaningful objective function are considered in this study, definitions of which are listed in the first row of Table II.2. They are all under maximization scheme and pursue the highest degree of correlation in general sense.

Let n_1 and n_2 be the sample sizes of the two subregions R_1 and R_2, leading to correlations ρ_1 and ρ_2. The first type of objective function is a weighted average of ρ_1^2 and ρ_2^2. Here the sample sizes are used as weights, and the squares of ρ_1 and ρ_2 are used to eliminate the sign effect (e.g., $\rho_1 > 0$, $\rho_2 < 0$). The aim of this objective function is to find a splitting that maximizes the overall correlation of X_1 and X_2 in the two subregions. The second type of objective function concerns the highest correlation among ρ_1 and ρ_2, where their absolute values are used to eliminate sign effect. The aim of this objective function is to identify a subgroup with the strongest correlation of X_1 and X_2. The third type of objective function focuses on the (absolute) difference between ρ_1 and ρ_2. This objective function aims to identify most distinguishable subgroups in correlations of X_1 and X_2.

Table II.2 Three types of objective function Ψ and corresponding stopping conditions

	Type 1	**Type 2**	**Type 3**
Objective function	$\Psi = \frac{n_1\rho_1^2 + n_2\rho_2^2}{n_1 + n_2}$	$\Psi = \max\{\lvert\rho_1\rvert, \lvert\rho_2\rvert\}$	$\Psi = \lvert\rho_1 - \rho_2\rvert$
Stopping condition	$\Psi - \rho^2 < \eta_1$	$\Psi - \lvert\rho\rvert < \eta_2$	$\Psi < \eta_3$

II.3.3 Controlling tree size

A correlation tree with complex structure (i.e., many branches and leaves) does not have easy interpretation. Thus, the size of tree should be controlled in tree building to favour small trees. In this study, two methods are used for this purpose. The first method is to specify some stopping conditions which signal the termination of splitting. The stopping conditions under the three types of objective function are given in the second row of Table II.2, where ρ is the correlation at the current node. For Type 1 objective function, splitting stops if the improvement (i.e., difference between the value of objective Ψ and ρ^2) is below a pre-specified threshold η_1. For Type 2 objective function, splitting stops if the improvement (i.e., difference between the value of objective Ψ and $|\rho|$) is below a pre-specified threshold η_2. For Type 3 objective function, splitting stops if the value of objective Ψ is below a pre-specified threshold η_3. The three thresholds η_1, η_2, η_3 are specified by the user, depending on their preference on tree size; larger values of them lead to smaller trees. Also, note that these thresholds are applied for differences between two correlations, and thus their values should be within $(0, 1)$. To provide an example, $\eta_1 = 0.1, \eta_2 = 0.1, \eta_3 = 0.25$ are used in the case study. The second method specifies the minimum sample size per node (e.g., ten samples based on literature [35]) and maximum tree depth (e.g., 3), i.e., the longest length from the root node to a terminal node, to achieve global control of tree size.

We want to mention that the goal of tree size control in correlation trees is slightly different from that in decision trees. In decision trees, controlling tree size is mainly to avoid the overfitting problem in addition to improving interpretability of the tree model. Overfitting means that the fitted model follows random errors, or noises, too closely; as a result, it has excellent performance in training, but may perform badly in prediction. To address this problem, two kinds of pruning have been used in the literature. Pre-pruning stops tree growing before it is fully grown by applying stopping conditions such as a threshold in improvement and minimum number of samples in leaf nodes. Post-pruning cuts back the fully grown tree to a sub-tree whose leaf nodes have the lowest cross-validation error. In contrast, correlation trees, designed for subgroup identification in terms of correlation instead of for prediction, are free of overfitting intrinsically. Thus, interpretability is the only motivation for controlling the size of a correlation tree, and the two pre-pruning methods used in this study are appropriate.

II.3.4 Time complexity

The learning process of correlation trees is the same as conventional decision trees, except for split variable selection. For linear correlation tree, the computational complexity is given by

$$\begin{cases} \mathcal{O}(nlogn), & for\ splitting\ point \\ \mathcal{O}(1), & for\ comparison \\ \mathcal{O}(n), & for\ partial\ correlation \end{cases}$$

where n is sample size at a node. For splitting point, tree considers all possible conceivable partition points to optimize an objective function whose sorting complexity is equivalent to that of CART algorithm [36], which dominates runtime in this step. It is obvious that time complexity of comparison among partial correlations over a split variable is $\mathcal{O}(1)$. As to complexity of partial correlation over the split variable Z_j, three components are involved: Least squares estimation on $X_1 \sim Z_j$, least squares estimation on $X_2 \sim Z_j$, and the Pearson's correlation between $\epsilon_{X_1 \sim Z_j}$ and $\epsilon_{X_2 \sim Z_j}$. The complexity of least squares estimation is $\mathcal{O}(n)$ since regression is performed with respect to a single predictor. The complexity of the Pearson's correlation is $\mathcal{O}(n)$ [37]. Thus, for k split variables, the global time complexity of linear correlation tree becomes $\mathcal{O}(kn + k + nlogn)$. In the case of nonlinear correlation tree, the complexity of partial correlation computation is tantamount to the computation of Spearman's rank correlation whose complexity is $\mathcal{O}(nlogn)$ [38]. Thus, the global time complexity of non-linear correlation tree becomes $\mathcal{O}(knlogn + k + nlogn)$.

II.4 Simulation study

A simulation study is done to validate the unbiasedness of the proposed split variable selection method. The basic idea is as follows: Generate data of two correlated variables X_1, X_2 and six different types of covariates $\{Z_1, Z_2, \dots, Z_6\}$ that are independent of X_1 and X_2; apply the proposed method to the simulated data and find the selected split variable; repeat this for a number of iterations and obtain the selection rate of each covariate (i.e., the percentage of iterations when that covariate is selected). Unbiased variable selection is expected to yield similar selection rates among all covariates regardless of their types and distributions. An alternative method for split variable selection is to optimize split variable and cutpoint simultaneously like what CART does, which is known to suffer from selection bias towards variables with more possible cutpoints. Here we call it CART-like method for convenience. Performance of this method under the three types of objective function given in Section II. 3.2 will also be assessed in the simulation to compare with the proposed method.

In the simulation, X_1 and X_2 follow a bivariate normal distribution with mean $\begin{bmatrix}3\\5\end{bmatrix}$ and covariance matrix $\begin{bmatrix}1.5 & 1.2\\1.2 & 2.1\end{bmatrix}$. Among the six covariates, three are continuous and three are categorical. The setting of their distributions is given in Table II.3. Two scenarios are considered: Case 1 where the Z's are independent of each other and Case 2 where some of them are dependent. Various distributions are generated in each scenario to reflect situations in practice. Specifically, in Case 1, Z_1 follows a uniform distribution in [1, 5], Z_2 follows an exponential distribution with mean 1, Z_3 follows a standard normal distribution, and Z_4, Z_5 and Z_6 follow a multinomial distribution with 2, 3, 6, respectively, categories of equal probabilities. In Case 2, Z_1, Z_2 and Z_5 follow the same distributions as in Case 1, Z_3 is a combination of Z_2 and a sample from standard normal distribution scaled by 0.2, Z_4 follows a Binomial distribution depending on Z_3, and Z_6 follows a six-category multinomial distribution depending on Z_5.

Covariates Z_3, Z_4 and Z_6 are correlated with other variables in Case 2. Generating data of Z_3 is straightforward by adding data from the scaled standard normal distribution to those of Z_2. Generating data of Z_4 and Z_6 which are categorical variables is not obvious. For Z_4, we rely on a procedure based on logistic regression as listed in Table II.3 to generate its samples. For Z_6, as Z_5 is also a categorical variable, a joint distribution of them is specified, and their samples are generated simultaneously from that distribution.

Table II.3 Distributions of covariates in the simulation study

	Case 1 (independent)	Case 2 (dependent)
Z_1	Uniform	Uniform
Z_2	Exponential	Exponential
Z_3	Normal	Z_2+0.2Normal
Z_4	Multinomial (2)	*Multinomial (2) depending on Z_3
Z_5	Multinomial (3)	Multinomial (3)
Z_6	Multinomial (6)	Multinomial (6) correlated with Z_5

* Procedure to generate this distribution:

Step1. Given β_0 and β_1, calculate the probability $P(Z_3) = \frac{exp(\beta_0+\beta_1 Z_3)}{1+exp(\beta_0+\beta_1 Z_3)}$.

Step 2. Generate a uniform random variable U in the interval [0, 1].

Step 3. If $P(Z_3) < U$, Z_4 is assigned to category 1; otherwise, category 2.

1000 iterations are carried out in the study, with 1000 samples generated in each iteration. Results on selection rates in building LCT and NCT are shown in Figures II.5 and II.6. In each Figure, the upper panel displays results in Case 1 and the lower panel displays those in Case 2. In each panel, “CART-like (Type 1)”, “CART-like (Type 2)”, and “CART-like (Type 3)” correspond to results of the CART-like method under the three types of objective function and “Proposed” refers to results of the proposed method. From Figure II.5, as expected, variables with more possible cutpoints are more likely to be selected in the CART-like method in both cases, regardless of the objective function used. To be specific, the three continuous variables Z_1, Z_2 and Z_3 have much higher selection rates than the three categorical variables Z_4, Z_5 and Z_6. Among the categorical variables, Z_6, with the highest number ($2^{6-1} - 1 = 31$) of cutpoints, has higher selection rates than Z_4 (1 cutpoint) and Z_5 (3 cutpoints). In contrast, the proposed method yields similar selection rates, around $1/6 = 0.166$, among all the covariates. The same patterns exist when NCT is built according to Figure II.6. These results validate that the proposed split variable selection method is unbiased.

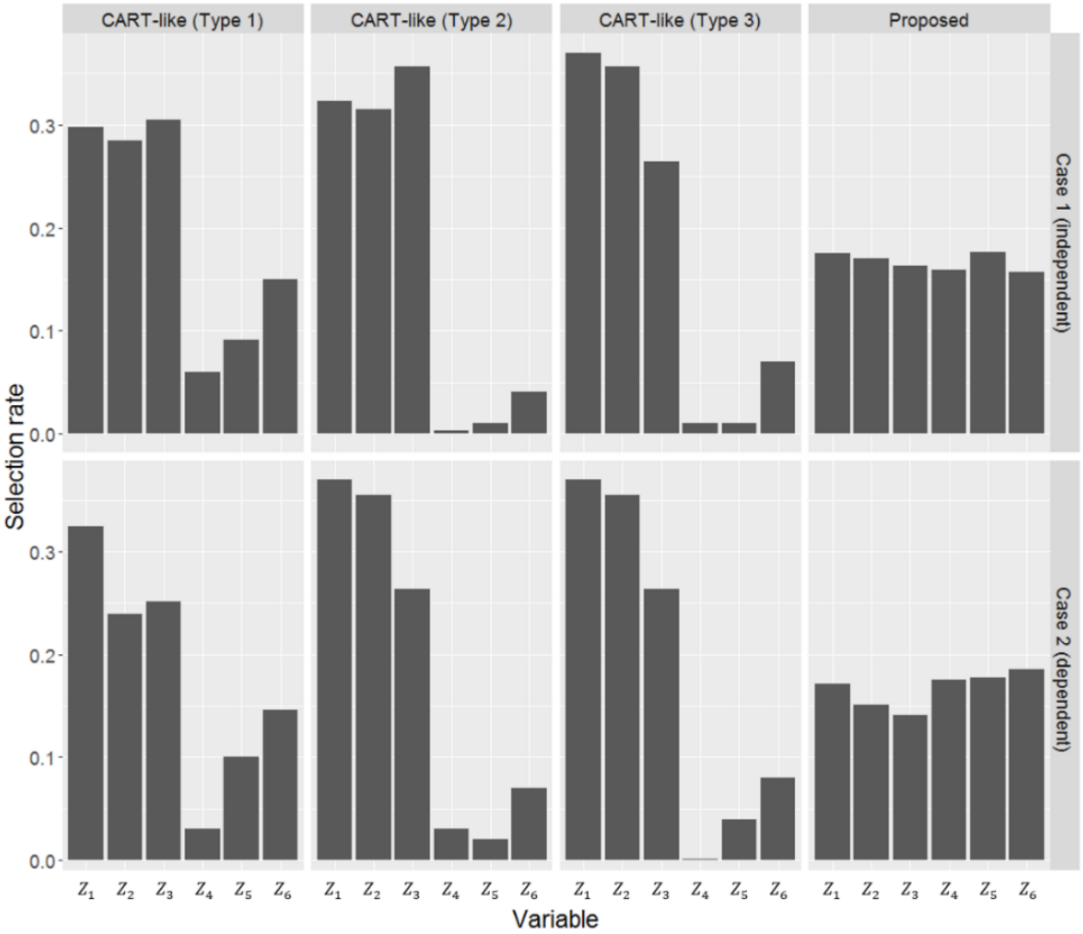


Figure II.5 Selection rates of all covariates in linear correlation tree

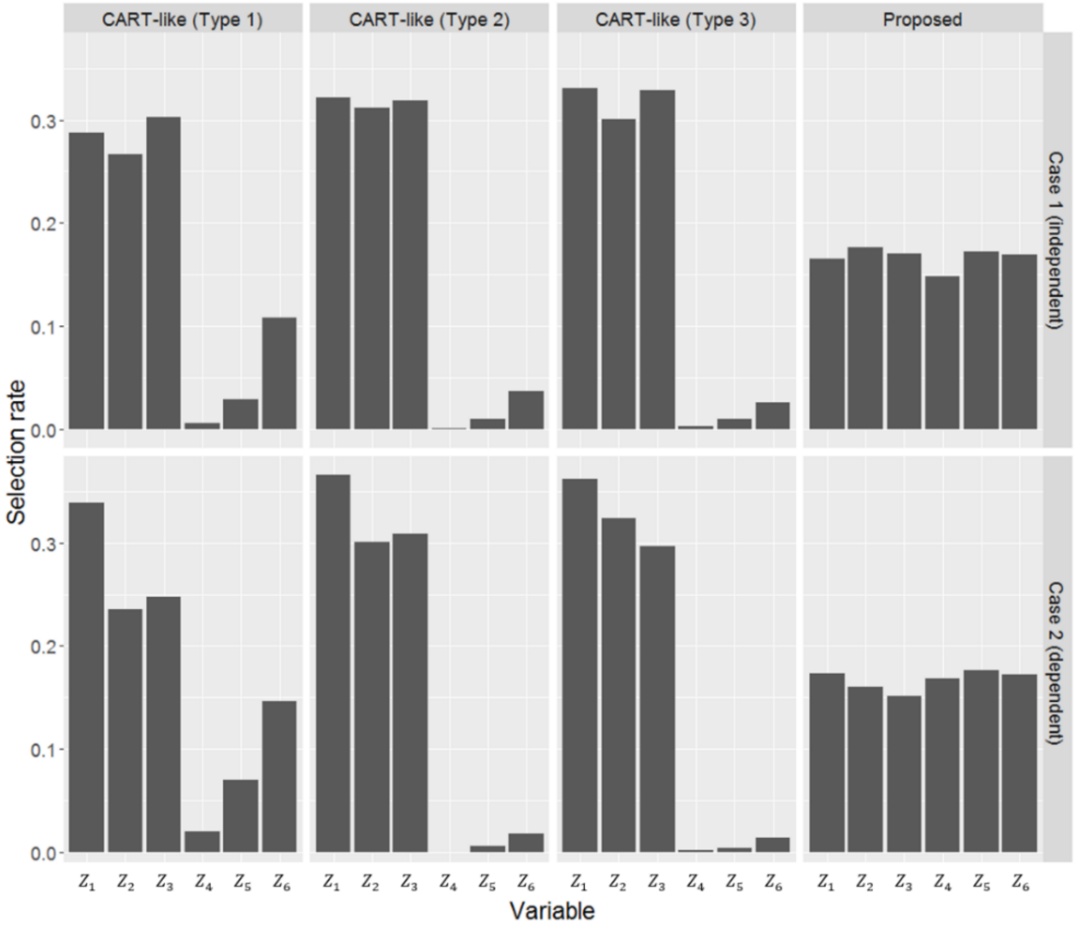


Figure II.6 Selection rates of all covariates in non-linear correlation tree

II.5 Application to real data

II.5.1 Data description

The proposed approach is applied to a dataset from an optical brain imaging study on risk decision-making [39]. In the study, each subject conducted a Balloon Analog Risk Task (BART) illustrated in Figure II. 7. The task contains 15 trials. In each trial, the subject sees virtual image of a balloon on the computer screen. He/she can choose to pump up the balloon or not under the risk that the balloon may explode. The trial ends up with two types of outcomes: *win* (the subject chooses to stop pumping and receives monetary reward proportional to the size of balloon) and *lose* (balloon explodes). Essentially, this task measures the degree to which subjects are willing to take a risk.

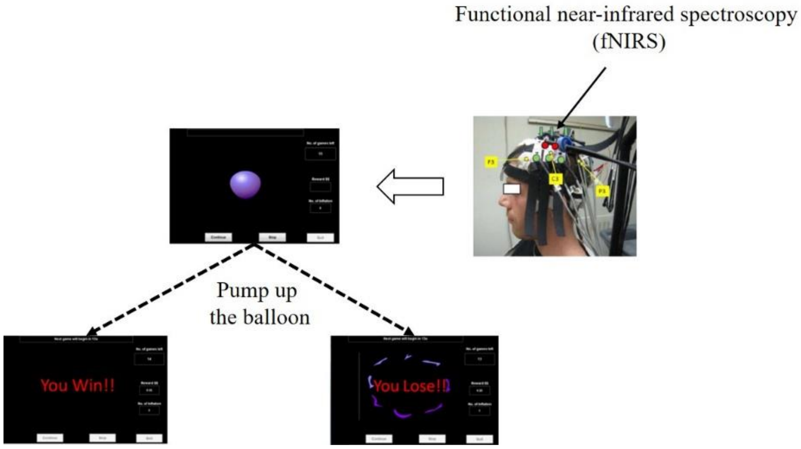


Figure II.7 Illustration of the optical brain imaging study

The correlation between brain activation and behaviour measure of subjects during the task is of interest. The brain activation is represented by ΔHbO, change in the concentration of oxygenated haemoglobin (iron-containing protein in red blood cells which carries oxygen from the respiratory organs to the rest of the body), and the behaviour measure is #pumps, the average number of pumps in each trial. As shown in Figure II.7, ΔHbO was measured by functional near-infrared spectroscopy (fNIRS), an optical brain imaging technique. A positive correlation between ΔHbO and #pumps implies a pattern of risk-taking, whereas a negative correlation implies a pattern of risk-aversion. The constructed correlation tree will help identify subgroups who differ in such patterns. There are five covariates that will serve as split variables in the tree: Gender, Education (four categories: High school or below, College, M.S., and Ph.D.), Age, systolic blood pressure (SBP), and diastolic blood pressure (DBP).

For simplicity in data analysis, data of each subject are aggregated in terms of outcomes of BART trials (i.e., win and lose). That is, the observation of a subject under the win/lose case is the average of his/her data over all the trials with win/lose outcomes. The sample size in the win and lose case is 91 and 92, respectively. Linear and nonlinear correlation trees will be constructed in the two cases separately.

Before building the trees, it is interesting to take a look at the brain-behaviour correlation in the whole space (i.e., using all data). Figure II.8 shows the scatter plots of ΔHbO and #pumps in the win and lose cases. In the win case, the two variables have a positive correlation (Pearson's = 0.513, Spearman's = 0.386), indicating a risk-taking tendency. In the lose case, they exhibit a negative correlation (Pearson's = –0.248, Spearman's = –0.224), indicating a risk-aversion tendency. The values of correlation measures are not high in both cases, which makes interpretation of the brain-behaviour relationship difficult. A reason for this lies in that large variation exists in the data and some samples look like outliers against the main trend.

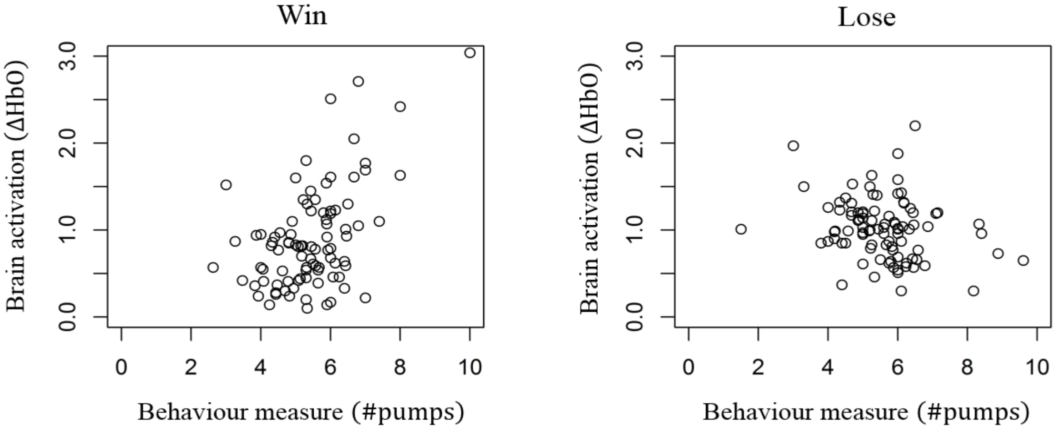


Figure II.8 Brain-behaviour correlation using all available data

II.5.2 Results of linear correlation trees

First, a linear correlation tree is constructed under each type of objective function. In the stopping conditions, the thresholds η_1, η_2 and η_3 are set to be 0.1, 0.1 and 0.25, respectively. The minimum sample size per node is set to be 10, and the maximum tree depth is set to be 3 levels where the root node is defined as level 0. This setting of control parameters is based on our purpose to identify as many subgroups as possible while achieving good interpretability.

Figures II.9 and II.10 show the constructed linear correlation trees in the win and lose cases. In the win case (Figure II.9), all LCTs select Age at the root node, while in the lose case (Figure II.10), all select Gender. The fact that the same variable is chosen in each case is not surprising, since the variable selection step is independent of the optimal cutpoint searching step; in other words, the first selected split variable is irrespective of the objective function used. Also, the selected variable in each case indicates that the trees are capable of identifying more significant covariates between Age and Gender. In fact, it is known that both Age and Gender play an important role in risk-decision making from the literature [39, 40]. However, existing studies do not point out relative importance of them. With an unbiased variable selection method, the proposed correlation tree can find the most significant covariate which provides additional useful information to understand the brain-behaviour relationship. Among the five split variables, three of them, Age, Gender, and SBP, appear in the trees. In particular, Age and Gender are involved in all the trees, suggesting that they are significant to define subgroups of brain-behaviour correlation. This is consistent with the convention in neural-correlates literature where age and/or gender effects are often considered.

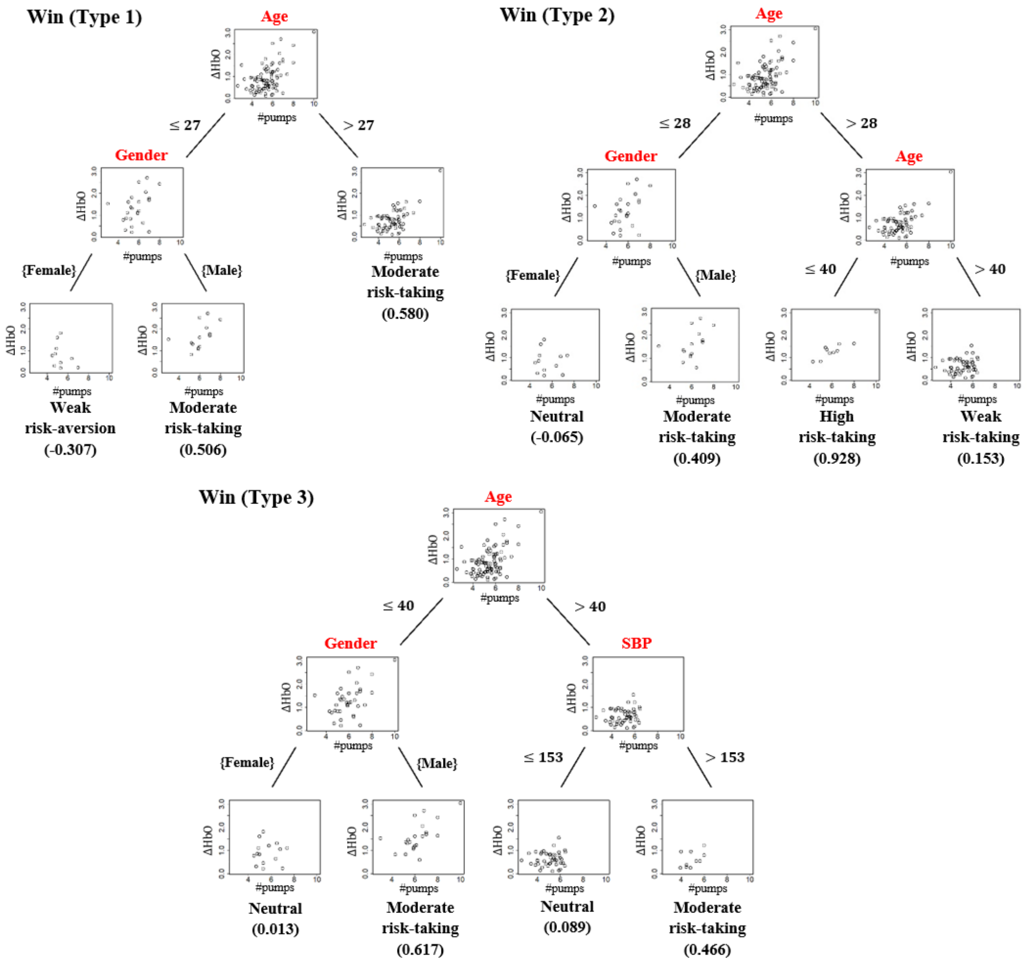


Figure II.9 The constructed linear correlation trees (LCTs) in the win case

In a general sense, correlation of two variables can be roughly categorized into four levels: neutral ($0 < \rho \leq 0.1$), weak ($0.1 < \rho \leq 0.4$), moderate ($0.4 < \rho \leq 0.7$) and high ($0.7 < \rho \leq 1.0$). Basically, in the win case, the identified subgroups exhibit a risk-taking pattern of different degrees from weak to high. For example, under Type 2 objective function, people older than 40 are weakly, males younger than 28 are moderately, and people older than 28 and younger than 40 are highly risk-taking. In contrast, subgroups in the lose case are inclined to avoid risk. For example, males older than 40 are highly risk-averse under all the three types of objective function (Pearson's = –0.884).

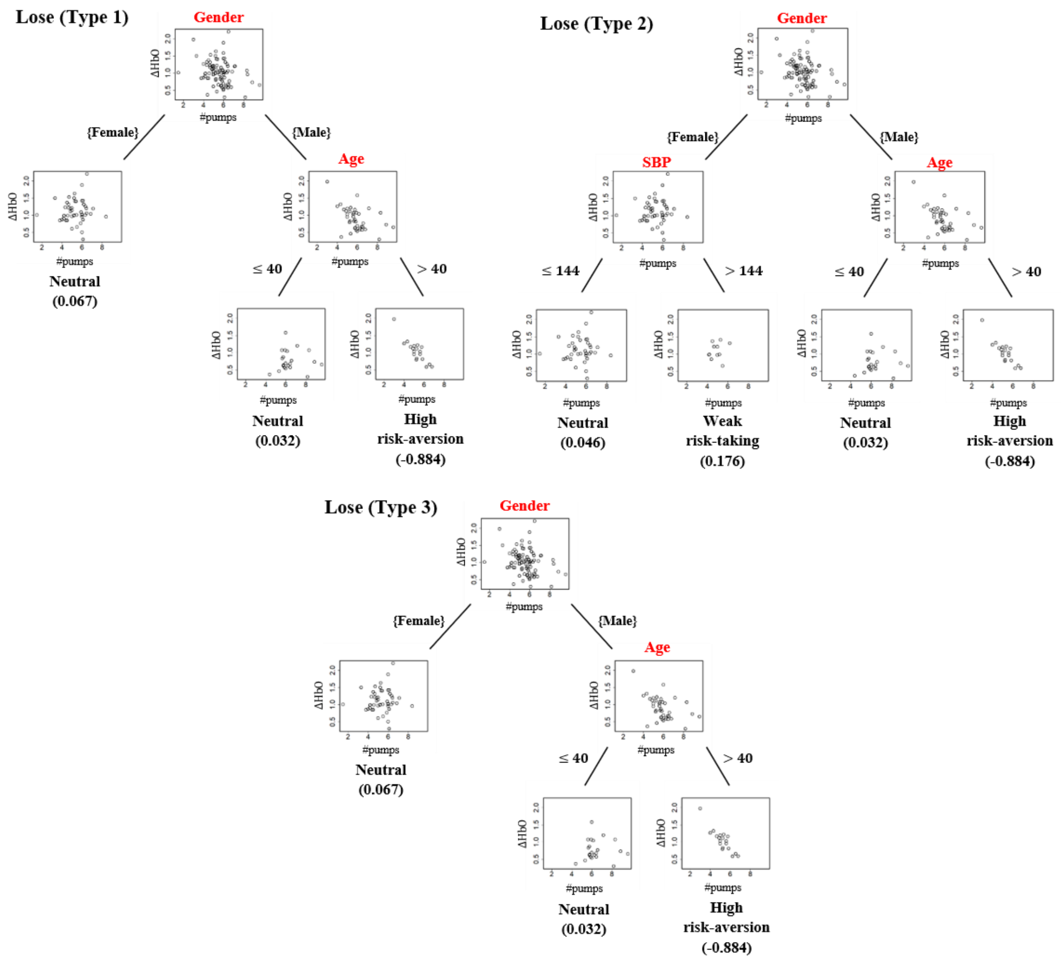


Figure II.10 The constructed linear correlation trees (LCTs) in the lose case

The objective function has an effect on the resulting subgroups. We will use the trees in the win case as examples to illustrate the effect. As mentioned in Section II.3.2, Type 1 objective function maximizes the overall squared correlations, and thus tends to produce similar subgroups (correlations of the three identified subgroups have magnitudes of 0.580, 0.506 and 0.307). Type 2 objective function focuses on the highest correlation, and thus it produces the highest risk-taking subgroup (with a correlation of 0.928). Type 3 objective function concerns most distinguishable subgroups, and thus it produces two pairs of subgroups in each of which one subgroup has almost zero correlation while the other has a large correlation.

The trees in the lose case are interesting in that they have almost identical subgroups, with neutral (Pearson's = 0.067, 0.032) and high (Pearson's = −0.884) correlations. Considering the population correlation (Pearson's = −0.224) shown in the right panel of Figure II.8, this indicates that the correlation tree successfully identifies hidden subgroups with strong correlation (males older than 40) or neutral correlation (males younger than 40 and all females), which lead to better interpretation of the brain-behaviour relationship. The similarity of results under the three types of objective function also implies that the identified subgroups may be determined by the intrinsic structure of the population, if any, and not sensitive to the objective function. To be specific, when a dominating subgroup (males older than 40 in this case) is obscured in the population, this subgroup will always be detected whichever objective function is used.

We want to point out that the resulting subgroups under a certain type of objective function may not be the most optimal in terms of the defined optimality. For example, in Figure II.9, Type 3 correlation tree is supposed to have the largest difference in correlation between a pair of subgroups. However, the actual largest difference in this tree is 0.604, which is smaller than 0.815 achieved by Type 1 correlation tree. This is because correlation trees, like decision trees, are subject to the inherent drawback of recursive partitioning, i.e., it is defined to achieve local optimum in each partitioning and cannot guarantee global optimum at the terminal nodes. Generally speaking, all the objective functions seek for subgroups in the direction of maximizing correlation. Therefore, their resulting correlation trees can be used in a complementary manner.

II.5.3 Results of non-linear correlation trees

With the same setting of control parameters as in building the LCTs, non-linear correlation trees are applied to the data. The resulting NCTs in the win case and lose case are shown in Figures II.11 and II.12. The identified subgroups in the NCTs are analogous to those in the LCTs. Four covariates are selected as split variables, including the three involved in LCTs, Age, Gender, SBP, and one additional, Education. One interesting result is that, according to Type 3 tree, there is a substantial difference in the brain-behaviour correlation of two education groups among people older than 40: those with college and more education are moderately risk-taking while those with less education do not have this trend. This indicates certain effect of education on human risk decision-making patterns, which is new to the literature.

However, the magnitudes of correlations in most NCTs are smaller than those in the LCTs. In particular, the differences are substantial in Type 1 and Type 2 trees in the lose case. This can be explained by the robustness of Spearman's rank correlation coefficient against outliers. Note that there are some potential outliers in the data, which have large influence on the correlation of each subgroup, especially when the sample size of the subgroup is small. They have likely caused the high correlations in the LCTs.

Spearman’s correlation coefficient is based on ranks rather than the original data, thus being able to alleviate the influence of outliers.

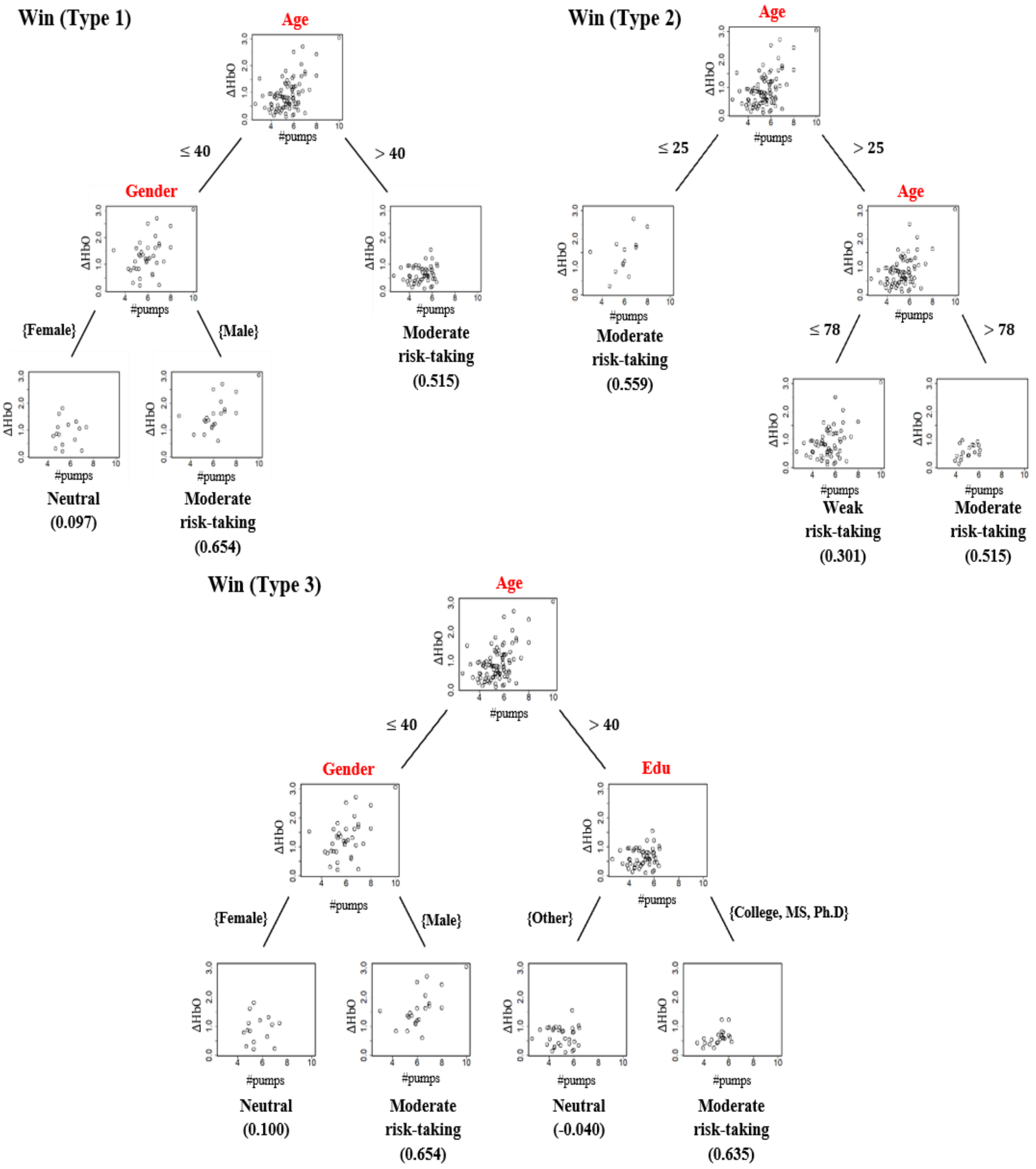
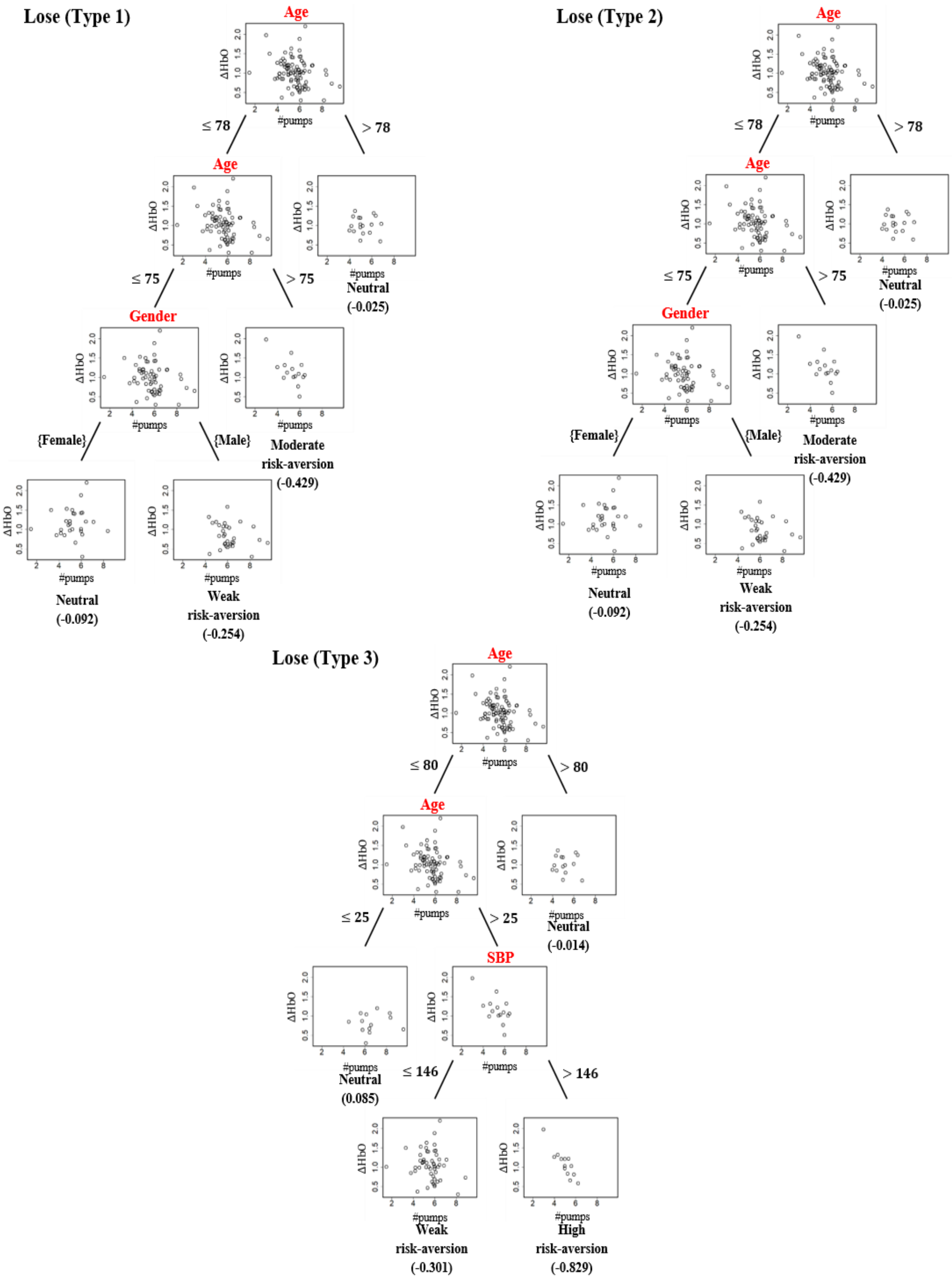


Figure II.11 The constructed non-linear correlation trees (NCTs) in the win case



Lose (Type 1)

Lose (Type 2)

Lose (Type 3)

Figure II.12 The constructed non-linear correlation trees (NCTs) in the lose case

II.6 Conclusion and discussion

In this study, we propose a recursive partitioning approach for subspace partitioning in terms of correlation of two variables. This approach can find important covariates associated with correlation as well as identify hidden subgroups of interest. The simulation study validates the unbiasedness of the method for split variable selection. Application of the proposed approach to a real dataset demonstrates that it can produce meaningful subgroups and significantly improve interpretability in correlation analysis.

This work is the first attempt to consider subspace partitioning with respect to correlation. The proposed correlation tree can identify subgroups of interest in an automatic and optimal way, which lays the foundation for a thorough understanding of the brain-behaviour relationship in neural correlates studies. Moreover, the proposed approach is not confined to neural correlates studies, but broadly applicable to other fields where correlation is a main concern.

Here are some general guidelines on the use of the correlation tree in practice. First, about the three types of objective function: If the goal is to obtain multiple subgroups with strong correlations, Type 1 objective function is suggested. Type 2 objective function is used to identify the subgroup with the strongest correlation, while Type 3 objective function is used to identify the most distinguishable subgroups. It is recommended that the resulting correlation trees under the three types of objective function be used in a complementary manner. Second, about linear vs. nonlinear correlation measures: nonlinear measures such as Spearman's rank correlation coefficient should be used if general correlation (not limited to linear form) is concerned and outliers are present.

Several interesting open problems remain in this study. First, missing values may exist in a dataset. In the literature on decision trees, some efforts have been made to solve this problem. For example, RPART [41] finds surrogate variables of a split variable and uses their data as substitute when the data of the split variable are missing. GUIDE [42] treats missing values as an additional category for a categorical variable, and sends all observations with missing values in the split variable to a subnode such that it leads to the greatest improvement on the objective function for a continuous variable. Those ideas could be adapted to handle missing values in correlation trees.

Second, outliers may also exist in practice, which can mask the true relationship in subgroups [43]. This study shows that the Pearson's is sensitive to, especially, univariate outliers (i.e., data points with unusual values in either X_1 or X_2), and the Spearman's is robust in this case. Unfortunately, however, some outliers in this study are indeed bivariate outliers (i.e., data points with unusual values in both X_1 and X_2), and the Spearman's is sensitive to such outliers because it rests on marginal distributions of X_1 and X_2, and thus does not capture the joint structure of the data [44]. Methods to deal with bivariate outliers in correlation trees will be investigated.

Third, in high-dimensional environments, tree-based approaches may not work very well. Powerful clustering algorithms such as hierarchical SOM [45] and polar SOM [46] may be used in this situation. To find clusters of interest in this study, correlation of X_1 and X_2 needs to be considered in the objective function of the chosen clustering algorithm. Moreover, as the identified clusters may not have easy interpretation as subgroups, some constraints need to be incorporated in clustering to guarantee interpretability. Finally, this study produces a number of correlation trees, as shown in Figure II.9- II.12, some of which may not be very different from each other. To eliminate the redundancy in tree representations, we can compare and integrate the trees by transforming each tree into a vectorial representation [47, 48].

CHAPTER III

ROBUST LOGISTIC REGRESSION TREE FOR SUBGROUP IDENTIFICATION IN HEALTHCARE OUTCOME MODELING[*]

Outcome data are routinely collected in healthcare practices and used for quality of care assessment and improvement. Logistic regression trees are a popular method for subgroup identification for binary outcome data. Outliers often exist in healthcare data, and many studies have addressed this problem with respect to model fitting in logistic regression. However, outlier problems are more complex in the context of tree models, as they involve subgroup identification in addition to model fitting. This study considers the outlier problem in logistic regression tree modeling of outcome data. It reveals the effects of outliers on split variable selection in identifying subgroups and proposes a method to construct logistic regression trees that are robust to outliers. The effectiveness of the proposed method and its advantages over alternatives are demonstrated in a simulation study and case studies.

III.1 Introduction

Measuring health care outcomes has become very important due to the increasing attention to the quality of care and a call for evidence-based practice [49]. Examples of

[*] Reprinted with permission from “Robust logistic regression tree for subgroup identification in healthcare outcome modeling” by Doowon Choi, and Li Zeng, IISE Transaction on Healthcare Systems Engineering, 10(3), Copyright © 2020, Taylor & Francis.

possible outcomes include patient mortality, adverse events, and hospital readmission, and all of these have been widely used as quality indicators for health services. For example, the Centers for Medicare and Medicaid Services (CMS) has undertaken a series of initiatives to reduce hospital readmission rates by penalizing hospitals with higher than expected rates and funding hospital-level improvements on reducing readmission [50]. Outcome data are routinely collected in healthcare practices [51, 52]. The analysis of such data aids healthcare customers in their decision-making and also increases the accountability of care providers and care quality.

Outcome data for a care process typically include outcome measures and related covariates such as personal characteristics, diagnoses, and treatment variables for each patient. The relationship between the outcome and covariates is often modeled in order to understand the effects of covariates on the outcome, identify important covariates, predict outcomes for future patients, or set baselines for monitoring care providers' performance in the long term. For binary outcomes (e.g., mortality) that are prevalent in healthcare, logistic regression is the most popular model due to its easy interpretability [53]. However, as the number of covariates increases, logistic regression becomes inadequate for modelling the complex relationship.

Logistic regression trees can overcome this limitation of logistic regression. The basic idea of logistic regression trees is to divide the population into a number of subgroups according to the covariates, so that a simple logistic regression can adequately explain the data for each subgroup. Such models have several merits over logistic regression. First, the use of simple logistic regression at each leaf node retains the easy interpretability of logistic regression. Second, the idea of subgroup-wise modeling is consistent with healthcare practices and easy for health professionals to understand. In addition, the nonlinear tree model fits the data more adequately than the linear logistic regression. For these reasons, logistic regression trees have been used in healthcare informatics to identify subgroups for binary outcome data [54-57]. The obtained subgroups will enable more accurate prediction of patient outcomes. Moreover, they lay a foundation for personalized medicine where optimal treatment is designed for each subgroup instead of the whole population. For example, the subgrouping of Type 2 diabetes patients can help the design of culturally suitable intervention programs to improve their self-care behaviors [54].

Outliers often exist in healthcare data. In general, outliers are defined as observations that deviate extremely from the bulk of the data [58]. The huge heterogeneity among individual patients, errors in collecting patient-related data, and overdispersion of healthcare data [59] are possible reasons for the presence of outliers. Outliers can have substantial effects in regression analysis, where they are accommodated at the expense of a poor fit for the majority of the data. Many studies have been conducted to address outlier problems in linear and logistic regression [60-63]. These studies can be roughly divided into two categories: outlier diagnostics/detection and robust estimation. Outlier diagnostics aims to pinpoint potential outliers that will be corrected or removed from the dataset before the formal analysis, while robust estimation attempts to restrict the influence of outliers in model fitting by, e.g., down-weighting them or using robust metric to outliters.

To the best of our knowledge, the outlier problem has not been thoroughly studied in the context of tree models. Tree models differ from linear and logistic regression in two respects: They involve identifying subgroups (i.e., splitting the covariate space) and model fitting for each subgroup. The effect of outliers on model fitting is similar to their effect in regression (i.e., they result in a poor fit), but the effect of outliers on subgroup identification is still unknown. It is believed that classification trees are relatively robust to outliers, as their splitting criteria, which are typically a function of proportions of classes, are not highly sensitive to outliers. However, regression trees, in which the splitting is based on the variance of the data, may be seriously affected. For example, due to outliers, a tree may fail to split a node that should be split or split a node that should not be split. This will result in misleading subgrouping in healthcare applications and affect the subsequent prediction or treatment based on it.

This study considers the outlier problem in logistic regression tree modeling of outcome data in healthcare. This problem is challenging in that it is difficult to quantify the effects of outliers on subgrouping. Moreover, the popular methods for addressing outliers in the regression literature, i.e., down-weighting and outlier detection, may not work for tree models. To conquer the problem, we first investigate, via a simulation, how outliers affect the identification of subgroups for building logistic regression trees. Based on the understanding gained from this investigation, we propose ideas to alleviate the outlier effect and make tree models robust to them. Our contributions are twofold:

- We reveal the effects of outliers on subgroup identification by focusing on how outliers affect split variable selection, which is critical for finding subgroups. Among the available algorithms for building logistic regression trees, we choose the MOdel-Based (MOB) recursive partitioning algorithm [64] as a benchmark for illustrating our idea.

- We propose a logistic regression tree that is robust to outliers. To demonstrate the effectiveness and advantage of the proposed method, we extend down-weighting and outlier detection to logistic regression trees and use them as reference in comparison. The case study validates that the proposed robust tree produces meaningful subgroups in presence of outliers.

The remainder of this paper is organized as follows. Section III.2 reviews the literature on logistic regression trees, the MOB algorithm, and methods to address outliers in logistic regression. Section III.3 uses a simulation to examine the effects of outliers on split variable selection. Section III.4 describes the proposed robust logistic regression tree and the two alternative methods, and Section III.5 compares the performance of the three methods through a simulation. Section III.6 applies the proposed method to two healthcare datasets. Finally, Section III.7 concludes the paper and discusses future research directions.

III.2 Literature review

III.2.1 Logistic regression trees

The first logistic regression tree was proposed by Chaudhuri *et al*. [65], who extend the SUPPORT (Smoothed and Unsmoothed Piecewise Polynomial Regression Trees) algorithm to binary responses. At each node, the probability that the response from a logistic regression will be 1 and a smoothed estimate of this probability using nearestneighbor averaging [66] are calculated for each observation. The difference between the two is called a “pseudo-residual”, and the observations are divided into two groups by the signs of their pseudo-residuals. Then, for each covariate variable, the two groups are compared to see whether there is a significant difference between their means and variances using a two-sample t-test and the Levene test [27]. The variable with the smallest p-value is chosen as the split variable, and the average of the two group means with respect to the selected split variable is taken as the cutpoint.

The LOTUS (Logistic Tree with Unbiased Selection) algorithm [67] adopts ideas that are similar to GUIDE (Generalized, Unbiased, Interaction Detection and Estimation) [28] for fitting logistic regression trees. Like GUIDE, LOTUS grows a tree by splitting each node in an unbiased fashion. It is known that covariates that allow more possible cutpoints are more likely to be selected as split variables when the splitting is based on minimizing the total sum of squared residuals or deviances at the two sub-nodes [68]. To nullify such a bias, LOTUS first selects the most significant unbiased split variable using a trend-adjusted chi-square test and then obtains the cutpoint by minimizing the sum of the deviances.

Landwehr *et al*. construct the Logistic Model Tree (LMT), which learns a logistic regression model at each node in an incremental manner (i.e., by boosting) [69]. The idea is that LMT does not link log odds with a linear predictor at each node as standard logistic regression does, but instead uses a target function called the “committee”, which is formed by combining many weak learners. For splitting, LMT relies on the information gain ratio used in C4.5 [70] and selects the covariate with the maximum gain ratio as the split variable. Lee and Jun improve the computational efficiency of LMT by adopting a least-angle regression in the boosting process [71].

More recently, a theoretical advancement in logistic regression trees has been achieved by the MOB algorithm. MOB is a unified framework for constructing trees with a parametric model that can be fitted using M-type estimators (e.g., a least squares estimator and a maximum likelihood estimator) at each terminal node. MOB selects the split variable based on the score function of the M-estimation. The key advantage of this algorithm lies in its integration of recursive partitioning and statistical model estimation/variable selection, which sets a rigorous theoretical foundation. To find the optimal split variable for each node, MOB examines the change in model parameters with respect to each covariate using a parameter instability test [72]. The test makes use of the full model scores and considers all possible changes, which is an improvement over other algorithms that only use partial information such as the sign of the pseudo-residual and ad-hoc approximations of possible change points [64]. We use a logistic regression tree built by the MOB algorithm (hereafter an "MOB tree") as our benchmark to study the effects of outliers.

III.2.2 MOB logistic regression tree

III.2.2.1 Basic concept

Let Y be the binary outcome measure, X the predictor in a logistic regression, and $\{Z_j: 1 \leq j \leq l\}$ the set of covariates. The predictor is used only for regression, while the covariates are used only for splitting the tree. By partitioning the covariate space, an MOB tree explores a piecewise logistic regression model $\{\mathcal{M}_b: Y \sim LG(X; \boldsymbol{\theta_b}), b = 1, \dots, \mathcal{B}\}$ that fits observations in each subgroup b better than a global model $\{\mathcal{M}: Y \sim LG(X; \boldsymbol{\theta})\}$, where $\boldsymbol{\theta}_b$ and $\boldsymbol{\theta}$ are respectively model parameters for the subgroup b and the entire population, $\mathcal{B}$ is the total number of subgroups, and "LG" is short for "logistic regression".

An example MOB tree is given in Figure III.1, where there are two split covariates (a patient's age and gender) and one predictor (the patient's risk score), and the response is the outcome of surgery (survival/death). At each terminal node, a logistic regression model $log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$ is fitted, where P is the probability of mortality after surgery and X is the patient's risk score. The model has two parameters: β_0 is the intercept, which represents the mortality rate of healthy patients (i.e., $X = 0$), and β_1 is the coefficient of X, which represents the effect of a patient's preoperative risk on the patient's mortality rate. The tree divides the patient population into three subgroups, depending on their gender and age, and the fitted logistic regression models for the subgroups have different parameter values.

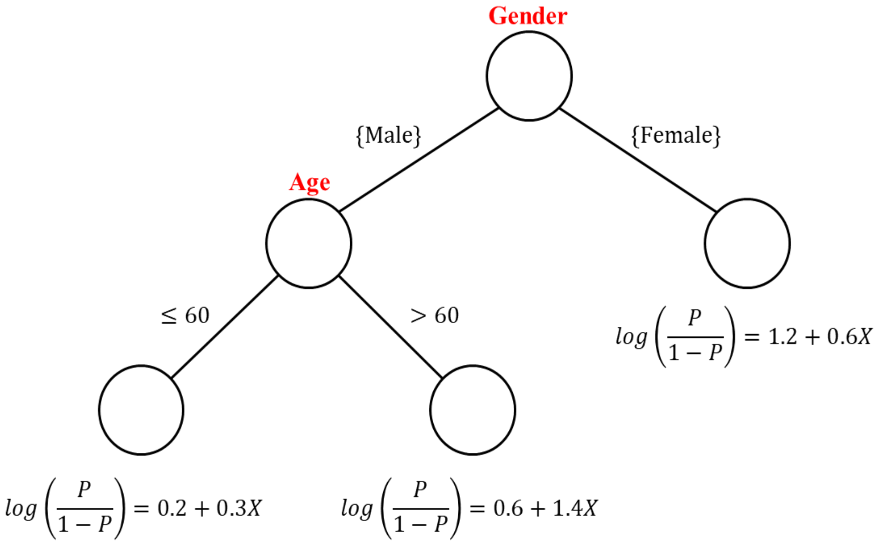

Gender
{Male}
{Female}
Age
≤ 60
> 60
$log\left(\frac{P}{1-P}\right) = 1.2 + 0.6X$
$log\left(\frac{P}{1-P}\right) = 0.2 + 0.3X$
$log\left(\frac{P}{1-P}\right) = 0.6 + 1.4X$

Figure III.1 An example MOB logistic regression tree

The MOB tree is constructed by recursively partitioning the covariate space and fitting a logistic regression for each of the resulting subgroups. Two steps are involved in identifying the subgroups: split variable selection and cutpoint estimation. Specifically, at each node, the best split variable among the covariates $\{Z_1, Z_2, \ldots, Z_l\}$ is selected, and then the optimal cutpoint for the selected split variable for forming subgroups is found. Details of the two steps are provided next.

III.2.2.2 Subgroup identification

The split variable at each node R is selected based on a parameter instability test, as illustrated in Figure III.2. The test is conducted for each covariate Z_j, $1 \leq j \leq l$. There are two options: (i) A global model with parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is adequate to explain the data at this node, and thus R should not be split by Z_j, as shown in the upper left panel of Figure III.2. In this case, the parameter is said to be *stable* with respect to Z_j. (ii) A global model is inadequate, and it is better to split R into two subgroups

$$R_1 = \{X | Z_j \leq c\}, R_2 = \{X | Z_j > c\}$$

using the cutpoint *c* and then fit two separate models to the subgroups with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, as shown in the lower left panel of Figure III.2. In this case, the parameter is said to be *unstable* with respect to Z_j. Based on this idea, whether node R is split depends on which of these two options (i.e., not split vs. split) is more plausible or, equivalently, the test of parameter instability (i.e., stable vs. unstable.). Accordingly, the covariate that has the strongest evidence for unstable parameter will be selected as the split variable.

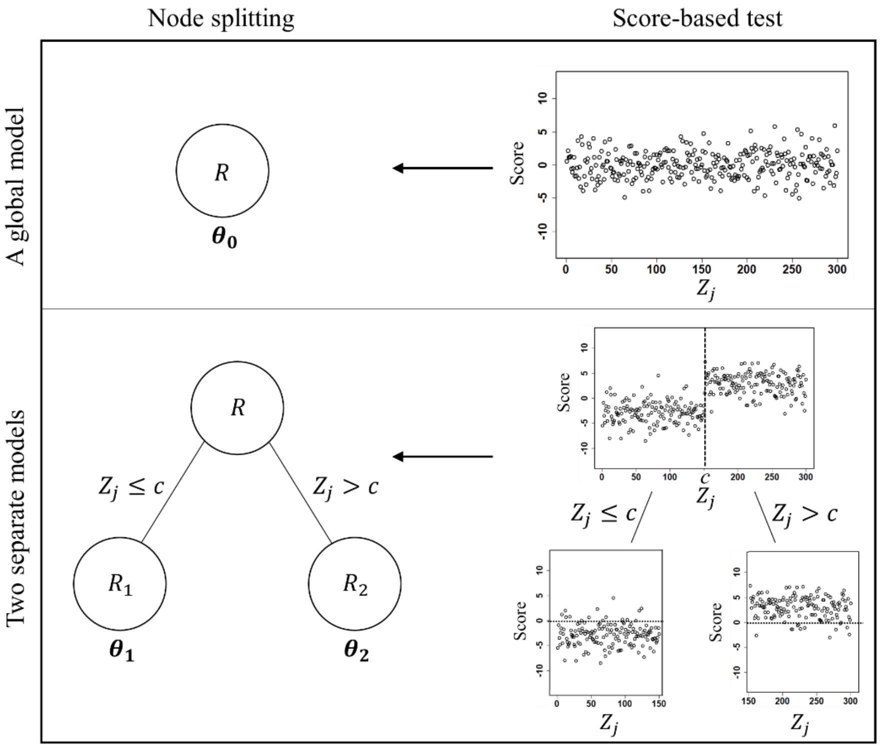


Figure III.2 Illustration of the parameter instability test for split variable selection

Formally, the test for parameter instability can be formulated as the following problem:

$$\begin{aligned} &H_0 : Y_i \sim LG(X_i; \boldsymbol{\theta}_0) \quad for\ i = 1, \dots, n \\ &H_1 : Y_i \sim \begin{cases} LG(X_i; \boldsymbol{\theta_1}) & if\ \boldsymbol{X}_i \in R_1 \\ LG(X_i; \boldsymbol{\theta_2}) & if\ \boldsymbol{X}_i \in R_2 \end{cases}, \end{aligned} \tag{III.1}$$

where n is the sample size of the data available at node R. This formulation is essentially a general form of change detection. Note that while regular change detection concerns change with respect to time, the change detection in Equation (III.1) concerns change with respect to the covariate variable Z_j; that is, the data are indexed by the order of Z_j, and the change point c for Z_j divides the data into two groups with different parameters. The p-value of the test indicates the strength of the alternative hypothesis; a smaller p-value means that it is more plausible to split the node. Thus, the covariate with the smallest p-value in the test will be selected as the split variable for this node.

The hypothesis test in Equation (III.1) is realized by a score-based test that is an adaptation of generalized M-fluctuation tests [73]. The rationale of the test is illustrated in the right half of Figure III.2. Given the data $\mathcal{D} = \{D_i = (Y_i, X_i), i = 1, \dots, n\}$ at node R, a global logistic regression model can be fitted by minimizing the negative log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D})$, yielding the maximum likelihood estimate

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \ell(\boldsymbol{\theta}; \mathcal{D}) \; = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \textstyle\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; D_i), \tag{III.2}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. The score function is defined as the partial derivatives of the objective function in the estimation

$$\boldsymbol{s}(\boldsymbol{\theta}; D_i) = \frac{\partial \ell(\boldsymbol{\theta}; D_i)}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ell(\boldsymbol{\theta}; D_i)}{\partial \theta_1}, \dots, \frac{\partial \ell(\boldsymbol{\theta}; D_i)}{\partial \theta_k}\right). \tag{III.3}$$

By the definition of the maximum likelihood estimate,

$$\textstyle\sum_{i=1}^{n} \boldsymbol{s}\big(\widehat{\boldsymbol{\theta}}; D_i\big) = \mathbf{0}. \tag{III.4}$$

That is, the score function evaluated at the maximum likelihood estimate, i.e., $\boldsymbol{s}\big(\widehat{\boldsymbol{\theta}}; D_i\big), i = 1, \dots, n$, has a zero mean. In other words, when a global model fits the data well (H_0 in Equation (III.1)), the scores $\boldsymbol{s}\big(\widehat{\boldsymbol{\theta}}; D_i\big)$ randomly fluctuate around zero, as shown in the upper right of Figure III.2. On the other hand, if the parameter changes with respect to Z_j (H_1 in Equation (III.1)), the scores will not fluctuate around zero but exhibit a certain pattern, e.g., most scores in subgroup R_1 are negative while most scores in subgroup R_2 are positive, as shown in the lower right of Figure III.2.

Based on this idea, the empirical cumulative score process over Z_j is constructed to capture the systematic pattern of the scores

$$\boldsymbol{W}_j(t) = \hat{\boldsymbol{J}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{s}\left(\hat{\boldsymbol{\theta}}; D_{(i|Z_j)}\right),\ 0 \leq t \leq 1, \qquad \text{(III.5)}$$

where $\hat{\boldsymbol{J}}$ is an estimate of the covariance matrix of the scores, $\lfloor nt \rfloor$ is the integer part of nt, and $D_{(i|Z_j)}$ is the observation corresponding to the ith smallest value of Z_j. That is, $\boldsymbol{s}\left(\hat{\boldsymbol{\theta}}; D_{(i|z_j)}\right)$ is the rearranged $\boldsymbol{s}(\hat{\boldsymbol{\theta}}; D_i)$ in Equation (III.3) according to the order of Z_j. Given n finite observations, t can take the values $\left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$, which represent the proportions of data up to the ith observation. Thus, $\sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{s}\left(\hat{\boldsymbol{\theta}}; D_{(i|z_j)}\right)$ is the overall lack of fit up to the nt^{th} observation. The estimate of the covariance matrix $\hat{\boldsymbol{J}}$ can be obtained by the outer product of the gradient (OPG) or the observed information matrix. The inverse square root of $\hat{\boldsymbol{J}}$ in Equation (III.5), i.e., $\hat{\boldsymbol{J}}^{-1/2}$, decorrelates the scores of the k parameters $\theta_1, \ldots, \theta_k$ so that we can inspect each parameter separately.

Under the null hypothesis in Equation (III.1), $\boldsymbol{W}_j(t)$ converges to a Brownian bridge $\boldsymbol{W}^0$ by the functional central limit theorem [61]. A test statistic $\mathcal{T}$ can be constructed by applying a scalar functional Λ to $\boldsymbol{W}_j(t)$. Equations (III.6) and (III.7) are two specific forms of $\mathcal{T}$, depending on the type of Z_j:

$$\mathcal{T} = \Lambda\left(\boldsymbol{W}_j(t)\right) = \max_{i=\underline{i},\ldots,\bar{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n}\right)^{-1} \left\| \boldsymbol{W}_j\left(\frac{i}{n}\right) \right\|_2^2 \text{ when } Z_j \text{ is a continuous covariate, (III.6)}$$

$$\mathcal{T} = \Lambda\left(\boldsymbol{W}_j(t)\right) = \sum_{q=1}^{Q} \frac{|I_q|}{n}^{-1} \left\| \Delta_{I_q} \boldsymbol{W}_j\left(\frac{i}{n}\right) \right\|_2^2 \text{ when } Z_j \text{ is a categorical covariate.} \quad \text{(III.7)}$$

The statistic in Equation (III.6) is the maximum of the squared L_2 norm of the cumulative score process scaled by its variance, $\left(\frac{i}{n} \cdot \frac{n-i}{n}\right)$, over the interval $[\underline{i}, \overline{\iota}]$, where $\underline{i}$ is the minimal segment size to guarantee the least powerful test for change detection and $\overline{\iota} = n - \underline{i}$. Typically, $\underline{i}$ includes 10% of the entire set of observations [74]. The limiting distribution for this statistic is the supremum of a squared, k-dimensional tied-down Bessel process [75]. The statistic in Equation (III.7) is the weighted sum of the squared L_2 norm of $\Delta_{I_q}\boldsymbol{W}_j\left(\frac{i}{n}\right)$, the increment of the cumulative score process over observations in category q (with associated indexes I_q). Its limiting distribution is a χ^2 distribution with $k(Q-1)$ degrees of freedom [76]. The test based on $\mathcal{T}$ is conducted for each covariate Z_j, and the one with the smallest p-value that is less than a pre-determined significance level α corrected for multiple (l) testings is selected as the split variable.

Once the split variable is determined, node *R* is split into two subgroups at the cutpoint *c*. This cutpoint is expected to produce the most distinguishable subgroups with respect to parameter heterogeneity. The cutpoint can be estimated by an exhaustive search of all conceivable cutpoints of the split variable for the one whose resulting subgroups yield the maximal reduction of the negative log-likelihood.

III.2.2.3 Tree size control

In decision trees, the tree size is usually controlled to avoid overfitting and enhance the interpretability of the fitted tree model. The same consideration applies in fitting logistic regression trees. Specifically, the MOB tree relies on two ways to control the tree size. First, it uses a pre-pruning technique that prevents the tree from fully growing by imposing a stopping criterion. The significance level α in the parameter instability test naturally serves as such a criterion; a smaller α leads to a smaller tree. Second, the minimum number of samples at each terminal node is also specified to further control tree size; the larger this number, the smaller the tree.

III.2.3 Outlier diagnostics and robust estimation in logistic regression

Several methods have been developed to detect outliers in logistic regression [62, 77-81]. The basic idea of those methods is to use residual analysis. Usually, residual plots against the predictor variable or the fitted probability of the response being 1 are drawn to identify outliers with large residuals. In order to measure how “large” a residual is, it is necessary to find an approximate distribution of the residuals.

Most studies on robust estimation of logistic regression impose down-weights on outliers to restrict their influence on model fitting. For example, Pregibon proposed resistant fitting methods that taper the deviance to limit the impact of extreme observations by using Huber’s weight function [82]. Stefanski *et al.* [83] and Kunsch *et al*. [84] adjusted the original score function to achieve bounded sensitivity, i.e., the maximum possible influence of a single observation. Morgenthaler replaced the L_2-norm in logistic regression with the L_1-norm, leading to a weighted score function [85]. Croux and Haesbroeck improved the Bianco and Yohai estimator by reducing the large leverage values using a proper weight [86]. Rather than down-weighting, Hobza *et al*. introduced a median estimator for the logistic regression [87]. Park and Liu [88] and Park and Konishi [89] proposed a robust logistic regression using regularization techniques.

It is worth noting that both outlier diagnostics and robust estimation rest on the assumption that all data are generated from a single model. In other words, the above methods will yield the most plausible single model fitting to the data in the presence of outliers. However, when heterogeneous subgroups exist, the effects of outliers become more complex, and direct application of these methods may no longer be effective. Thus, an investigation of the outlier effects in the logistic regression tree context is warranted.

III.3 Effects of outliers on split variable selection

III.3.1 Setup for the simulation

For convenience, we consider a simple scenario where there is a predictor X and a single covariate T that represents time. In the example given in Figure III.1, this means that we are concerned with whether the parameters of the logistic regression model $log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X$ experience any change over time. Two scenarios are simulated, as illustrated in Figure III.3: one in which the parameters of the logistic regression remain constant over time and one in which the parameters change at a certain time point t^*; hereafter we call these the *no-change case* and the *change case*, respectively. For each scenario, a dataset is first generated from the assumed model, with random outliers added. Then the parameter instability test with the statistic in Equation (III.6) is conducted to decide whether to split the data by the covariate T. Consequently, the effects of outliers are obtained by assessing the performance of the parameter instability test in finding the true underlying scenario for the data.

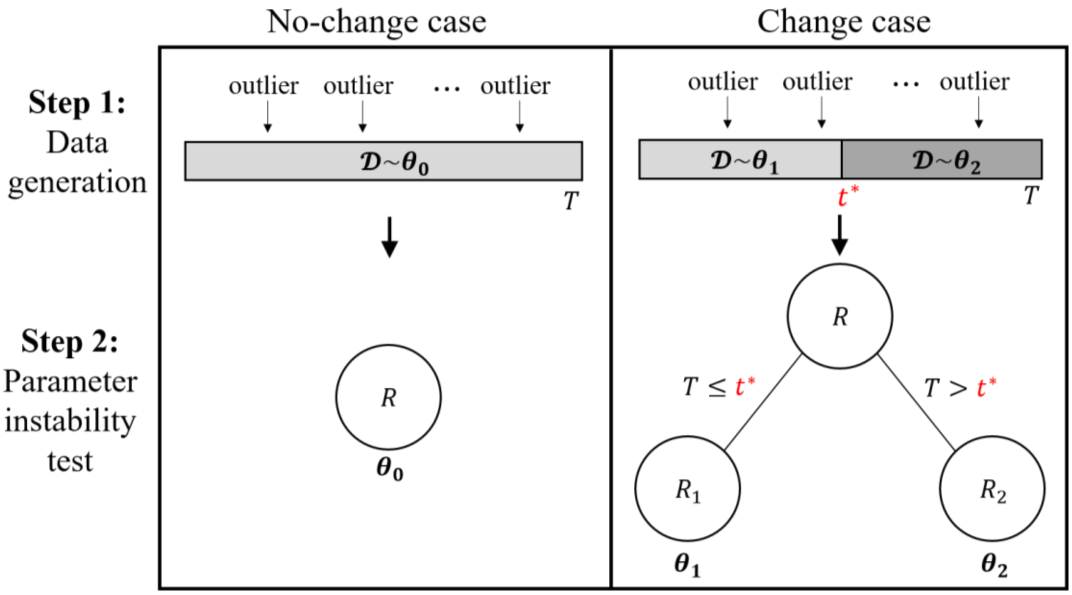


Figure III.3 The two scenarios considered in the simulation study and the corresponding node splitting in the parameter instability test

Specifically, X follows a discrete uniform distribution over $[1, 9]$. In the no-change case, the parameter $\boldsymbol{\theta}_0 = (\beta_0, \beta_1)$ takes the value $(-4.5, 0.775)$. In the change case, the pre-change parameter $\boldsymbol{\theta}_1$ is the same as $\boldsymbol{\theta}_0$, while the post-change parameter $\boldsymbol{\theta}_2$ takes a different value. Assuming the change occurs in β_0, six different values, $\beta_0 = \{-3.6, -3.9 - 4.2, -4.8, -5.1, -5.4\}$, are considered for $\boldsymbol{\theta}_2$ to reflect a wide range of possible changes. For convenience, the change point t^* is assumed to be the middle point of the simulated period. The outliers are generated by forcing y to be 0 when the probability that $y = 1$ is high and forcing it to be 1 when that probability is low, which is equivalent to generating y from parameters with the opposite signs [90]. The severity of the outlying is controlled by the proportion of outliers in the data; five different proportions, $\{2\%, 4\%,\ 6\%,\ 8\%,\ 10\%\}$, are considered. For each proportion, the added outliers are distributed uniformly.

Different performance measures are used for the two scenarios. As shown in Figure III.3, in the no-change case, the parameter instability test is correct if it fails to reject the null hypothesis, which is that the data are not split. Thus, the performance of the test in this case can be measured by the probability of a Type I error or the *false splitting rate*, i.e., how likely it will incorrectly decide to split. In contrast, in the change case, the parameter instability test is correct if it rejects the null hypothesis, meaning that the data are split by T. Therefore, the performance of the test in this case can be measured by the probability of a Type II error or the *miss splitting rate*, i.e., how likely it will incorrectly decide not to split. To assess the performance in each scenario, 1000 runs were simulated, and the percentage of runs with false splitting or miss splitting was calculated. The results are summarized in the following two subsections.

III.3.2 Effect of outliers in the no-change case

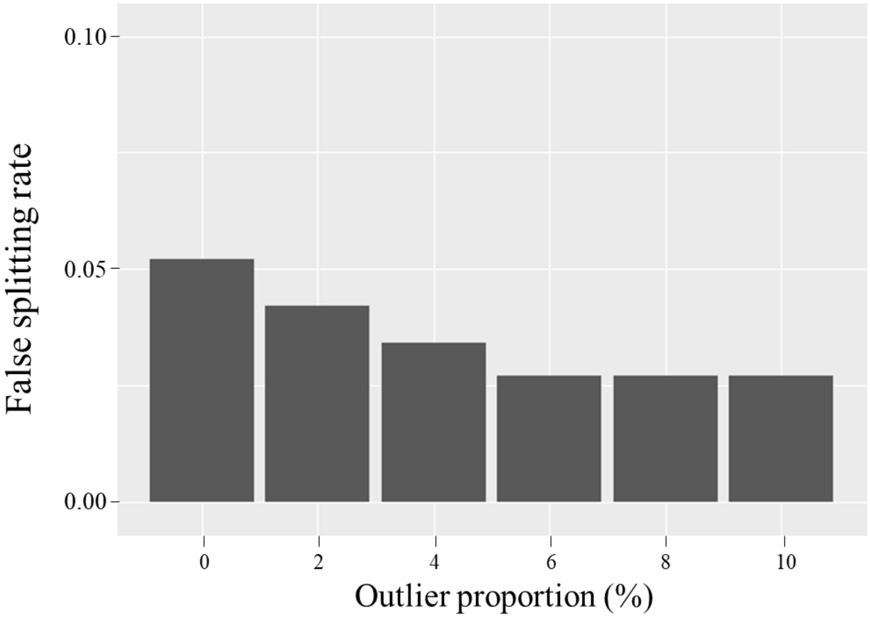


Figure III.4 Results for the false splitting rate in the no-change case

Figure III.4 shows the results for the no-change case. The false splitting rate in the absence of outliers (i.e., the outlier proportion is 0%) is around 0.05, which is consistent with the specified significance level $\alpha = 0.05$ in parameter instability tests. However, seemingly counterintuitively, as more outliers are present in the data, the false splitting rate decreases somewhat, which implies that it is less likely that a node is falsely split. We can conclude that in the no-change case, outliers are beneficial insofar as they can help decrease the false splitting rate.

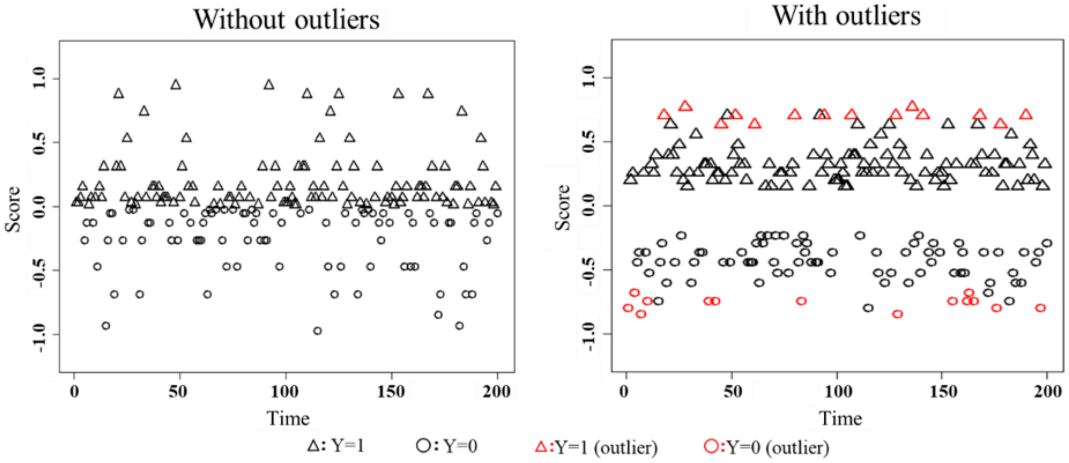


Figure III.5 A simulated example to illustrate the outlier effect in the no-change case

To provide an intuitive understanding of this effect of outliers, Figure III.5 gives a simulated example of scores in the parameter instability test in the no-change case, where observations with $y = 1$ and $y = 0$ are denoted by triangles and circles, respectively, and outliers are marked in red. Without outliers, the scores are randomly distributed around zero, indicating no parameter change over time. With outliers, the scores still randomly fluctuate around zero, but those near the center move away from the zero line while those on the edges move closer to the zero line. As a result, the scores become more similar to each other and form a flatter pattern, which presents stronger evidence of parameter stability. In other words, the test is more likely to decide not to split the node, leading to a lower false splitting rate.

The pattern in the right panel of Figure III.5 can be explained by the masking and swamping effects of outliers [91, 92]. It is known that when outliers exist, the model fitting tries to chase such atypical observations in order to fit all data well, and the resulting fitted line is likely to be close to outliers while deviating from the majority of the data. Thus, the outliers tend to have smaller or similar residuals compared to other observations and appear to be normal; in other words, the existence of outliers is masked. On the other hand, normal observations are swamped, that is, they have larger or similar residuals compared to the outliers. The masking and swamping effects of outliers produce the pattern of residuals in the right panel of Figure III.5. Since the scores are a function of residuals, they exhibit the same pattern. When the proportion of outliers in the data becomes larger, the effect of outliers becomes stronger, and thus the false splitting rate decreases.

III.3.3 Effect of outliers in the change case

Figure III.6 shows the results for the change case. The six panels represent different magnitudes of change. In each panel, the higher the proportion of outliers, the higher the miss splitting rate. We can conclude that in this case, outliers increase the miss splitting rate, meaning that the fitted tree tends to have fewer branches than needed (i.e., it is a smaller tree). It can also be seen in Figure III.6 that the miss splitting rate decreases when the magnitude of change increases, which is expected, as a larger change is easier to detect.

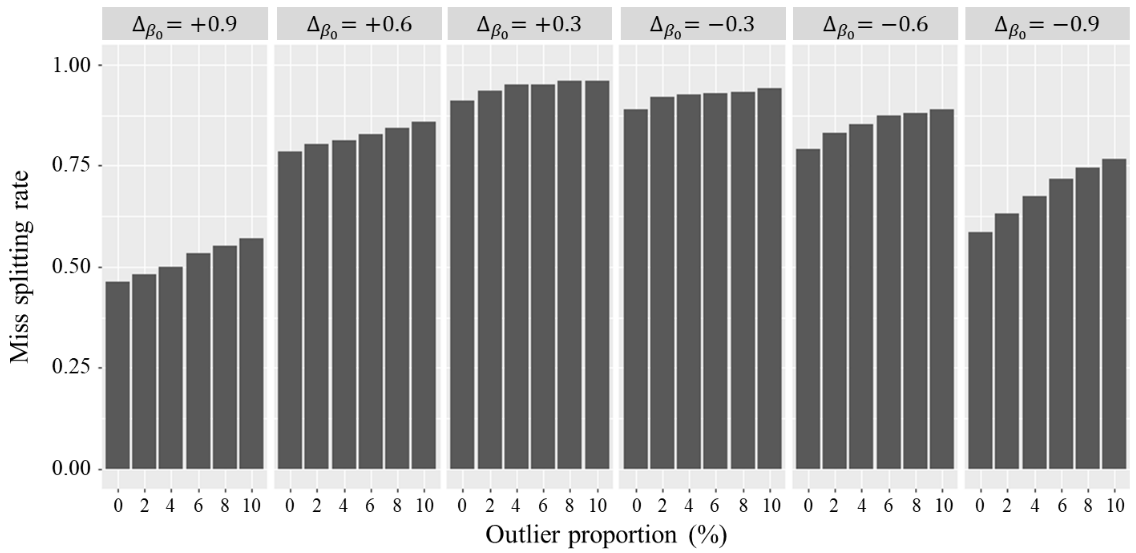


Figure III.6 Results for the miss splitting rate in the change case

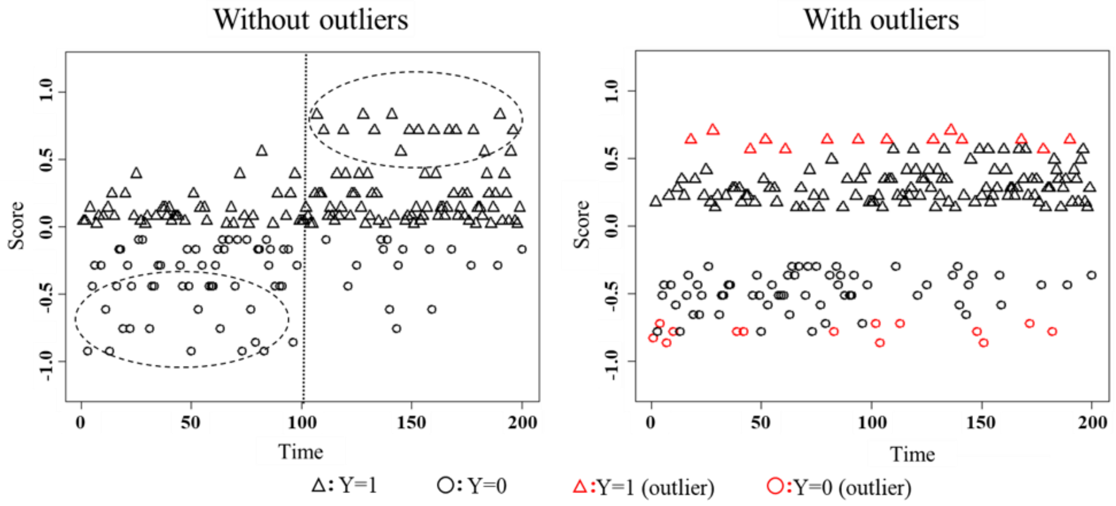


Figure III.7 A simulated example to illustrate the outlier effect in the change case

This effect of outliers can also be explained by the masking and swamping effects of outliers. Figure III.7 provides a simulated example of scores to illustrate this. When the data contain no outliers, there is a clear structural pattern indicating a change in the parameter occurring around $T = 100$. Note that the points with large scores (circled) before and after the change point are the main contributors to the evidence of change. In this study, we call such points “*change indicator points*”. However, as a result of the masking and swamping effects, the scores for those change indicator points become smaller, while those for normal observations become larger, forming a flatter pattern that presents weaker evidence of change, as shown in the right panel of Figure III.7.

Note that the increase in the miss splitting rate caused by outliers may be much more severe in practice when the number of covariates l is large. As the parameter instability test for each covariate is conducted with a Bonferroni-corrected significance level of α/l, a larger l will lead to a smaller significance level and thus a higher miss splitting rate. The increase due to outliers in this case may make the miss splitting rate unacceptably high, even under large changes.

III.4 Robust logistic regression tree

The simulation study in Section III.3 shows that outliers slightly reduce the false splitting rate and considerably increase the miss splitting rate in split variable selection. We consider three methods to alleviate the increase in the miss splitting rate and make the subgroup identification more robust to outliers. The first two methods extend the conventional ideas for addressing outliers in logistic regression, i.e., down-weighting and outlier detection, to logistic regression trees. The third is our proposed method, which modifies the conventional outlier detection method.

III.4.1 Down-weighting

As mentioned in Section III.2.3, the idea behind down-weighting is to restrict the influence of outliers on model estimation by imposing down weights on them. The application of this idea in logistic regression is called “robust logistic regression”. To build a down-weighted logistic regression tree, a robust logistic regression is first fitted at node R, and then the scores of the fitted model are used in the parameter instability test for split variable selection. Among the available robust logistic regression methods, we chose the one proposed by Pregibon [70] for this study, as it uses Huber’s robust M-estimation, whose score function can be directly used in the parameter instability test.

This method imposes weights on observations with large deviances, which are potential outliers, by using Huber’s loss function in the parameter estimation. The robust estimator for logistic regression is

$$\widehat{\boldsymbol{\beta}}_{robust} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} h(X_i)q\left(\frac{d_i(Y_i,\boldsymbol{X}_i;\boldsymbol{\beta})}{h(\boldsymbol{X}_i)}\right), \tag{III.8}$$

where $q(u)$ is a tapering function, $h(X_i)$ is a factor that handles the leverage of each observation, i.e., that point’s degree of deviation from other points in the predictor space, and d_i is the ith observation’s deviance. If $q(u) = u$ and $h(X_i) \equiv 1$ in Equation (III.8), this estimator reduces to the standard maximum likelihood estimator. When $h(X_i) \equiv 1$, the tapering function is expressed as

$$q(d_i) = \begin{cases} d_i & d_i \le H \\ 2(d_i H)^{1/2} - H & otherwise \end{cases}, \tag{III.9}$$

where H is a predetermined threshold for deviance that often takes the value 1.345^2 [93]. The derivative of the tapering function is

$$w(d_i) = \frac{\partial q(d_i)}{\partial d_i} = \begin{cases} 1 & d_i \le H \\ (H/d_i)^{1/2} & otherwise \end{cases}. \tag{III.10}$$

The weight w in Equation (III.10) is known as "Huber's weight". It controls the undue influence of outliers by imposing a weight smaller than 1 that is inversely proportional to the square root of the deviance of a potential outlier.

Once the robust parameter estimates are obtained, the robust version of the score function and the covariance matrix of scores in Equation (III.5) can be found by

$$\boldsymbol{s}_{robust} = \sum_{i=1}^{n} w_i \boldsymbol{X}_i \left\{ Y_i - g^{-1}\left(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_{robust}\right)\right\}, \tag{III.11}$$

$$\widehat{\boldsymbol{J}}_{robust} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - g^{-1}\left(\boldsymbol{X}_i^T \widehat{\boldsymbol{\beta}}_{robust}\right)\right)^2 (w_i \boldsymbol{X}_i)(w_i \boldsymbol{X}_i)^T, \tag{III.12}$$

where $\boldsymbol{X}_i = [1\ X_i]^T$. The derivation is given in Appendix A.1. In these equations, g is the logit link function and $\widehat{\boldsymbol{J}}_{robust}$ is the OPG estimator of the covariance matrix [52]. The cumulative score process can be obtained by plugging Equations (III.11) and (III.12) into Equation (III.5), and then the parameter instability test statistic in Equation (III.6) or (III.7) can be calculated.

III.4.2 Outlier detection

The idea of outlier detection is to identify the outliers in the data at node R and perform the parameter instability test after removing them. As mentioned in Section III.2.3, outlier detection is usually made through residual analysis. However, it is difficult to define the residual for logistic regression. The conventional definition, i.e., $Y_i - \hat{p}_i$, is not adequate due to the binary nature of the response variable [67]. Pearson's chi-square residual [50, 65, 68], deviance residual [69], and binned residual [51] have been studied, but they rely on approximation of the residual distribution under special assumptions.

In this study, we use the Bayesian residual proposed by Albert and Chib [94] to detect outliers. The Bayesian residual is defined based on the latent-variable logistic regression [95]. This model is equivalent to the standard logistic regression. Table III.1 compares the two models. In the standard model, Φ denotes the logistic cumulative distribution function. In the latent-variable model, the latent variable ξ is a linear model of the predictor with random error ε, which follows a t distribution with 8 degrees of freedom to approximate the logistic link function. Y is assigned 1 if $\xi > 0$, and 0 otherwise.

Table III.1 Comparison of the two logistic regression formulations

Standard logistic regression	Latent-variable logistic regression
$Y \sim Bernoulli(p)$	$\xi = \boldsymbol{X}^T\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim t(8)$
$p(Y=1) = \Phi(\boldsymbol{X}^T\boldsymbol{\beta})$	$Y = \begin{cases} 1 & if\ \xi > 0 \\ 0 & if\ \xi \leq 0 \end{cases}$

The advantage of the latent-variable model is that by introducing a linear regression into the logistic regression, the well-defined residual analysis in linear regression can be extended to logistic regression. Specifically, the latent residual ε can be used to detect outliers. As ξ is a latent variable, ε is estimated using the Bayesian method. In the Bayesian framework, the residual ε_i of each observation is a random variable, so the posterior distribution of the residual $P(\varepsilon_i|\mathcal{D})$ needs to be obtained. The estimation procedure for $P(\varepsilon_i|\mathcal{D})$ is given in Appendix A.2. Further, we can define the outlying probability, i.e., the probability of an observation being an outlier, as

$$p_i^{out} = P\left(|\varepsilon_i| > \widetilde{K} \middle| \mathcal{D}\right), \tag{III.13}$$

where $\widetilde{K}$ is a constant. Letting $\widetilde{K}$ be a high percentile point of the prior of ε_i (i.e., a t distribution with 8 degrees of freedom), the outlying probability quantifies how far the ith observation deviates from the prior. For example, when $\widetilde{K}$ is the 97.5th percentile of the prior, observations with an outlying probability exceeding 0.05 will be taken as outliers.

III.4.3 The proposed method

The underlying assumption of the above two methods is that all data at a node come from a single logistic regression model. Accordingly, these methods select outliers by applying a pre-specified threshold (Huber's constant H in Equation (III.10) and the percentile point K in Equation (III.13)) for deviance or residual, and then assign smaller weights to or altogether remove them. This works when the data truly come from a single model, i.e., the no-change case. However, when there is parameter change in the data, the "outliers" selected by these methods may include not only the real outliers, but also non-outliers, including change indicator points, as shown in Figure III.7. As a result, down-weighting or removing the selected outliers may weaken the evidence of change or even make the change undetectable. Therefore, a robust logistic regression tree should work on real outliers while keeping the change indicator points unaffected.

Based on this understanding, we propose a method that modifies the original outlier detection method in Section III.4.2. Our modifications are threefold. First, a new metric that is more effective than the outlying probability in Equation (III.13) is defined to measure the degree of outlying. Second, a decision rule is established for the proposed outlying metric to determine potential outliers. Finally, intrinsic extreme points of the data are recovered from the selected outliers, and then the parameter instability test is conducted. The algorithm for the proposed method is summarized in Appendix A.3.

III.4.3.1 The proposed outlying metric

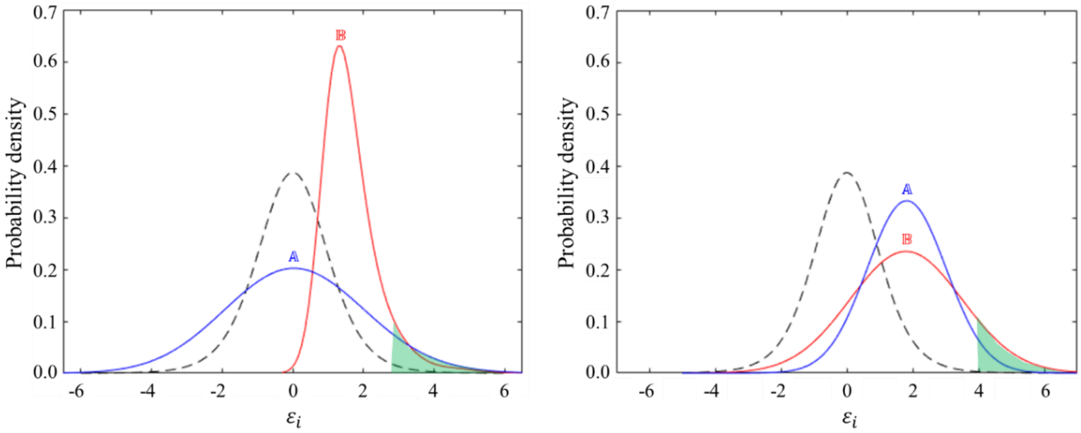


Figure III.8 Example of posterior latent residual distributions with similar outlying probabilities (left) and similar posterior means (right)

The new outlying metric is defined as

$$M_i = p_i^{out} \times |E[\varepsilon_i]|, \tag{III.14}$$

where $|E[\varepsilon_i]|$ is the absolute posterior mean of the latent residual. The motivation for combining the outlying probability p_i^{out} and $|E[\varepsilon_i]|$ is illustrated in Figure III.8. In each panel of Figure III.8, the dashed curve denotes the prior of ε_i as the reference, and the two solid curves are posterior distributions. The outlying probability for each posterior is represented by the shadow area under the curve. In the left panel, the two posteriors have similar outlying probabilities but different means. It is obvious that posterior $\mathbb{B}$ indicates more serious outlying than posterior $\mathbb{A}$, as its mean is farther away from zero. In the right panel, the two posteriors have similar means but different outlying probabilities, and posterior $\mathbb{B}$ indicates more serious outlying due to its higher outlying probability. The two examples suggest that the outlying probability or posterior mean alone is not adequate to distinguish outliers of different degrees, and the proposed metric in Equation (III.14), which combines them, shows a better ability to represent outliers.

III.4.3.2 The proposed decision rule

The original outlier detection method uses a constant threshold, i.e., $\widetilde{K}$, to decide on outliers. This type of decision rule will not work for the proposed metric in Equation (III.14), as it is hard to find the metric's distributional information. Moreover, as previously noted, a constant threshold, which is based on the single model assumption, may mistake non-outliers as outliers. In addition, determining the threshold value is challenging, as the appropriate setting depends on the specific scenario (e.g., no-change, small change, large change), which is unknown in practice.

In this Section, we propose a decision rule for the proposed metric that can solve the above problems. The idea is to find the group of observations that deviate most from the others. Specifically, we first sort the values of the proposed metric $\{M_i, i = 1, \dots, n\}$ in descending order and calculate the adjacent differences between the sorted values $\{M_{(i)}, i = 1, \dots, n\}$, where $M_{(1)} \geq M_{(2)} \geq \cdots \geq M_{(n)}$:

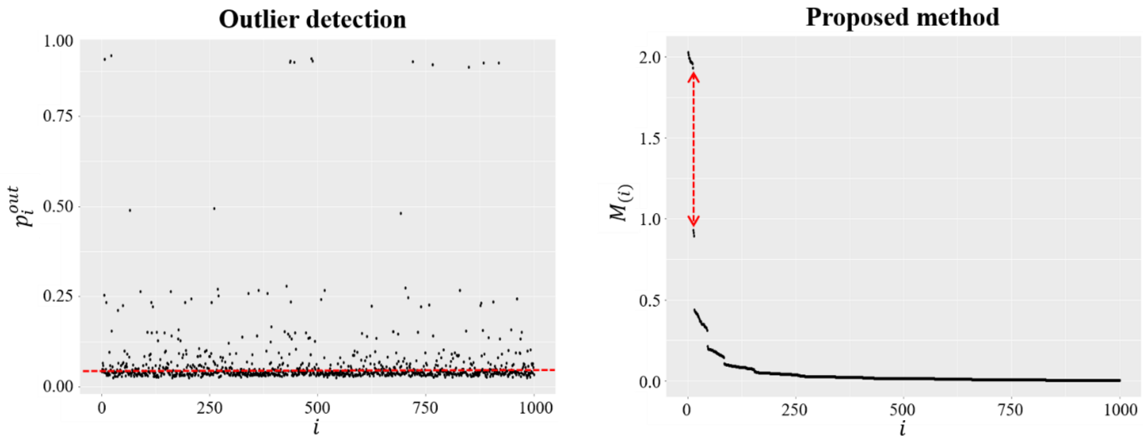
$$Chasm_i = M_{(i)} - M_{(i+1)}. \qquad \text{(III.15)}$$

Then we find the location of the largest adjacent difference:

$$i^* = \max_{1 \leq i \leq n-1} Chasm_i.$$

This means that the biggest gap among the sorted metric values occurs between the group $\{M_{(1)}, \dots, M_{(i^*)}\}$ and the remaining group $\{M_{(i^*+1)}, \dots, M_{(n-1)}\}$. Thus, the original observations corresponding to $\{M_{(1)}, \dots, M_{(i^*)}\}$ are taken as potential outliers. Considering that the Bayesian residual takes a positive or negative sign depending on the value of Y (1 or 0), this procedure is applied separately for each value of Y.

The proposed decision rule is intuitive and convenient to use, as it does not require any distributional information. It is designed to select observations with the highest degree of outlying, which are most likely the real outliers, and thus minimize the chance of mistaking change indicator points as outliers. Moreover, it is less sensitive to the specification of $\widetilde{K}$, as the outliers are determined not directly by the value of $\widetilde{K}$ but rather by the relative differences of the metric values.


Outlier detection

Proposed method

Figure III.9 Comparison of the outlier detection method and the proposed method

A simulation example is presented in Figure III.9 to illustrate the difference between the proposed method and the original outlier detection method. The simulated data contain a change and 12 outliers. The dashed line in the left panel denotes the decision threshold $p_i^{out} = 0.05$, with $K = 2.55$ calibrated to have a false splitting rate of 0.05 in the corresponding no-change case, while the arrow in the right panel denotes the biggest chasm found through the proposed method. Observations above the dashed line in the left panel and those above the biggest chasm in the right panel are the selected sets of potential outliers. The proposed method obtains 12 outliers, which are all the real outliers that are simulated. In contrast, the original outlier detection method produces many more outliers, which include the actual outliers but also many change indicator points.

III.4.3.3 Parameter instability test with recovery of intrinsic extreme points

In general, any real dataset intrinsically contains some extreme points due to the stochastic nature of the data generation process [96]. In the no-change case, those points are important for guaranteeing accurate estimation, while in the change case, those points play the role of change indicators and thus are critical for ensuring the good performance of change detection. Unfortunately, outlier detection methods are in general not able to distinguish between such intrinsic extreme points and outliers. That means that the selected potential outliers resulting from the proposed decision rule in Section III.4.3.2 may still include some intrinsic extreme points that should be recovered back to the data.

Let r be the number of intrinsic extreme points to be recovered from the selected set of potential outliers. A simple idea is to specify a value for r, randomly select r observations from the potential outliers, and return them to the normal samples. Then the parameter instability test is conducted using the updated normal samples. However, this overlooks the uncertainty in the number of intrinsic extreme points and the randomness in sampling. To take those into account, a range of values is considered for r, and multiple samplings are conducted to select intrinsic extreme points from the outlier set. The overall statistic in the parameter instability test will integrate all these possibilities.

The proposed procedure to conduct the parameter instability test with recovery of intrinsic extreme points is illustrated in Figure III.10. It is assumed that $L \leq r \leq U$, where L is the minimum possible number and U is the maximum possible number of intrinsic extreme points. Without any prior knowledge, L is set to 0 and U is the total number of selected outliers. Under each value of r, m trials are conducted, in each of which r samples are randomly drawn from the outlier set and returned to the normal data, and the statistic $\mathcal{T}$ in the parameter instability test is calculated using the updated normal data. Then the average of the m statistics, $\mathcal{T}_r^{ave}$, is obtained. This will produce a series of statistics $\mathcal{T}_L^{ave}, \mathcal{T}_{L+1}^{ave}, \ldots, \mathcal{T}_U^{ave}$. To maximize the chance of detecting a change, the overall statistic is the maximal average:

$$\mathcal{T}_{overall} = \max_r \{ \mathcal{T}_r^{ave}, r = L, L+1, \ldots, U \}. \qquad \text{(III.16)}$$

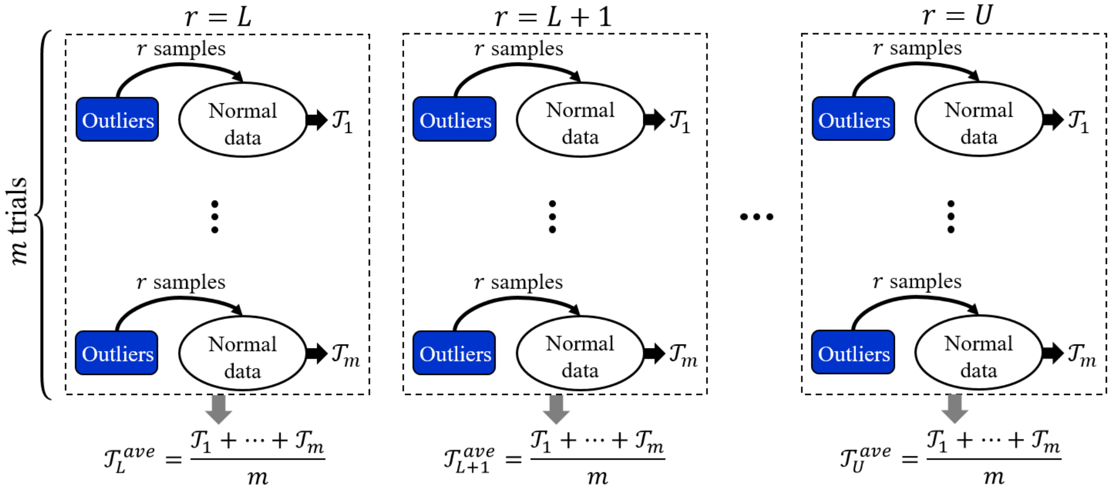

$r = L$
$r = L + 1$
$r = U$
m trials
r samples
Outliers
Normal data
$\mathcal{T}_1$
$\mathcal{T}_m$
$\mathcal{T}_L^{ave} = \frac{\mathcal{T}_1 + \cdots + \mathcal{T}_m}{m}$
$\mathcal{T}_{L+1}^{ave} = \frac{\mathcal{T}_1 + \cdots + \mathcal{T}_m}{m}$
$\mathcal{T}_U^{ave} = \frac{\mathcal{T}_1 + \cdots + \mathcal{T}_m}{m}$

Figure III.10 Illustration of the proposed procedure to conduct the parameter instability test with recovery of intrinsic extreme points

III.5 Performance comparison

The performances of the methods described in Section III.4 are compared in this Section based on a simulation. It is interesting to see the differences that can be attributed to the recovery of intrinsic extreme points in the proposed method, so a total of four methods are compared: (i) down-weighting, (ii) outlier detection, (iii) the proposed method without recovery of intrinsic extreme points, and (iv) the proposed method with recovery. The original parameter instability test, i.e., using all data without addressing outliers, will be used as a reference in the comparison.

The setup for the simulation is the same as in Section III.3. That is, two scenarios, a no-change case and a change case, are considered, and the false splitting rate and miss splitting rate are used as performance measures. In down-weighting, the threshold H is set to be 1.345^2. In outlier detection, the threshold $\widetilde{K}$ is set to be 2.55 so that the false splitting rate in the corresponding no-change case is 0.05. In the proposed method, the value of $\widetilde{K}$ is simply 2.306, the 97.5th percentile of the prior distribution, for convenience, and the number m of samplings for recovering extreme intrinsic points is 300. In the Bayesian posterior sampling, the number of samples is 10000 with 4000 burn-ins, which means that the first 4000 samples are discarded from the 10000 samples generated in each sampling.

The results of the performance comparison in the no-change case are given in Figure III.11. In each plot, the blue bar denotes the original parameter instability test and the red bar denotes the corresponding method. All methods show a similar false splitting rate when there are no outliers and a higher false splitting rate than the original test when outliers are present. This is because they attempt to alleviate the effect of outliers, which reduces the benefit of outliers in the false splitting rate described in Section III.3.2. One advantage of the three outlier-detection-based methods is that their false splitting rates are around the specified significance level 0.05 over all of the outlier proportions, making the parameter instability test quite stable and free of outlier influence. The reason is that these methods are able to identify most of the outliers and eliminate them from the data. Down-weighting exhibits slightly lower false splitting rates because it does not get rid of outliers, so that the benefit of outliers in the false splitting rate is partially retained.

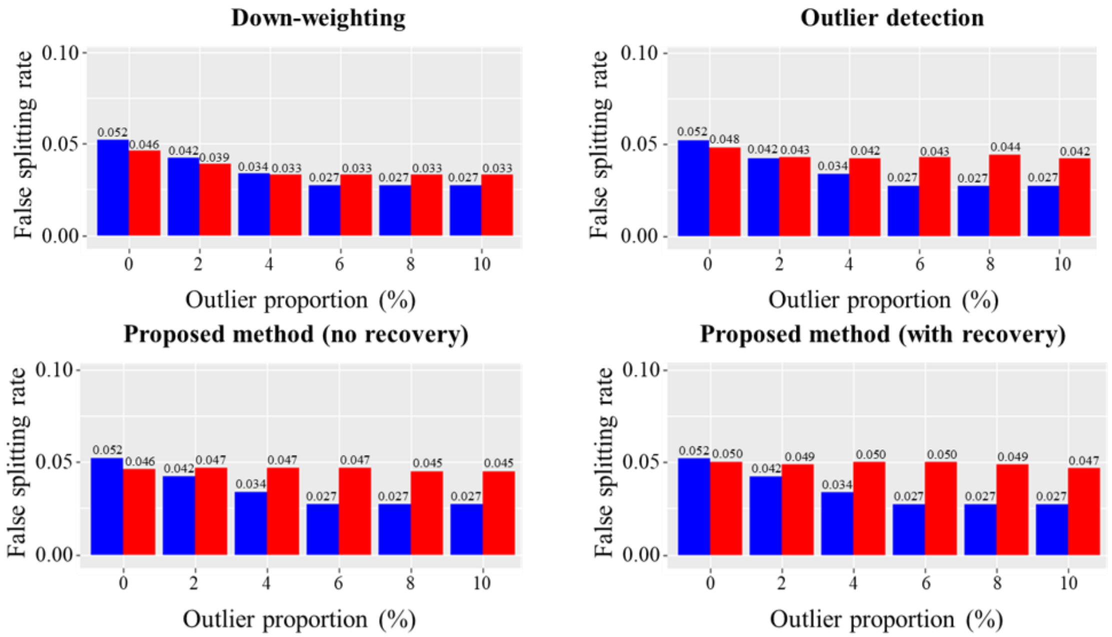


Figure III.11 Results for the false splitting rate of the four methods in the no-change case, with the blue bar in each plot denoting the original parameter instability test as a reference

Figure III.12 shows the performance of the four methods in the change case. Outlier detection yields a substantially higher miss splitting rate than the original test and the other methods in most cases because it removes change indicator points that are mistaken as outliers. Down-weighting has a larger miss splitting rate than the original test when no outliers exist, as it reduces the contribution of change indicator points by down-weighting them. In the presence of outliers, it has a similar or slightly lower miss splitting rate due to the restricted influence of outliers. The proposed method, with or without recovery of intrinsic extreme points, performs best in all cases because it preserves evidence of change while removing outliers. Regarding the two versions, the one without recovery has a similar miss splitting rate to the original test in the absence of outliers over all outlier proportions, implying that the proposed method can always successfully isolate real outliers from the data. The recovery of intrinsic extreme points further improves this performance.

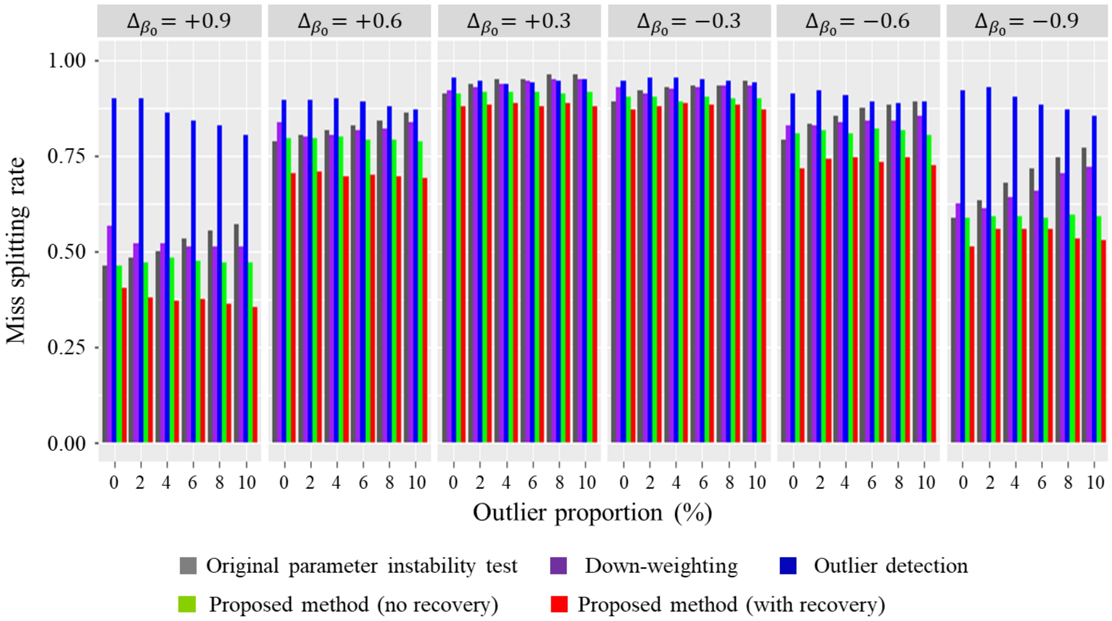


Figure III.12 Results for the miss splitting rate of the four methods in the change case

Another point that deserves mention is the trend of the performances under different magnitudes of change. Down-weighting has a similar miss splitting rate to that of the original test when the change is small (i.e., in the center panels), and a smaller rate when the change is large. The reason is probably that there are more change indicator points under a large change, and thus the weighting effect on the contribution of change indicator points becomes weaker. Like down-weighting, outlier detection performs similarly to the original test under small changes, but its performance under large changes is much worse because it removes most change indicator points from the data. The proposed method has a smaller miss splitting rate under all changes, and its advantage becomes more salient as the change becomes larger.

III.6 Case studies

This Section applies the methods to two healthcare datasets. The first dataset concerns hospital readmissions of chronic obstructive pulmonary disease (COPD) patients, and the second dataset concerns the mortality of patients in cardiac surgery. A logistic regression tree is built for each dataset using the five methods (regular MOB and the four methods addressing outliers) studied in Section III.5, and their results are compared.

III.6.1 Application to COPD data

COPD is the fourth leading cause of death in the world [97]. The dataset was collected during 2009–2012 and contains records for 420 patients [98]. The outcome measure is the *readmission of a patient within 30 days* after discharge, and 36 covariates are available, including patient baseline characteristics (e.g., demographics, comorbidities, and habitual behaviors), treatment variables (e.g., steroid usage, antibiotics), blood test results (e.g., hemoglobin, red blood cell distribution width), and history of health service utilization (e.g., the number of emergency room visits and hospitalizations).

When a single predictor for the outcome of interest is not specified, we first identify the important variables that affect the outcome among the pool of covariates. Since many covariates in this dataset are categorical variables, group LASSO is a suitable variable selection method. Seven variables are selected: *ER visits* (the number of emergency room visits), *Hospitalization* (the number of hospital stays), *RDW* (red blood cell distribution width), *Age*, *LAMA* (whether the patient was treated with long-acting muscarinic antagonists), *Alcohol* (whether or not the patient uses alcohol), and *Antibiotic* (whether the patient was treated with antibiotics). The predominant one, *ER visits*, is used as the predictor for logistic regression at each node, and other variables serve as split variables in the tree. In constructing the tree, the significance level for the parameter instability test is set to 0.05, the minimum sample size per node is set to 150, and the parameter settings of the proposed method are the same as in Section III.5.

Figure III.13 shows the logistic regression trees constructed by the different methods. The two versions of the proposed method produce the same result, so only one tree is displayed here. All of the trees have a simple structure, with only one splitting. The selected split variable is *Hospitalization* for regular MOB, down-weighting, and the proposed method. According to the experience of medical professionals in this field, the number of hospital stays is an important indicator for patient readmission. So the subgrouping based on *Hospitalization* is reasonable. The outlier detection method splits the patients by *RDW*, which is not appropriate. These results validate that the proposed method is able to produce the correct subgrouping

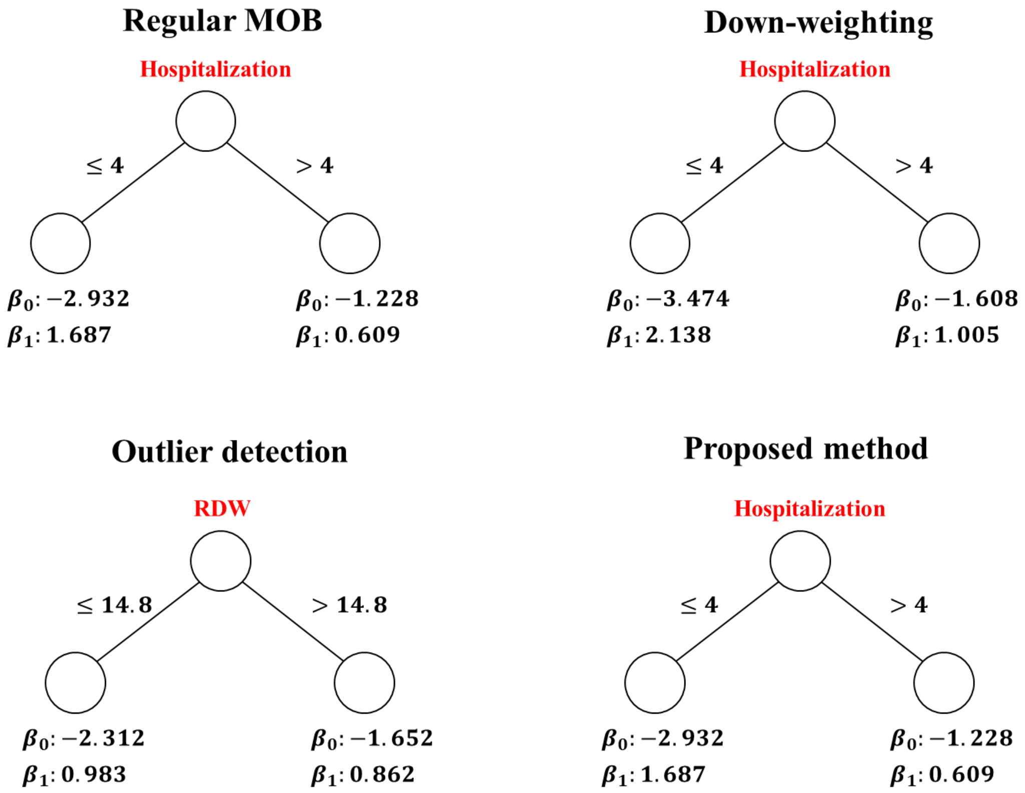


Figure III.13 The logistic regression trees for the COPD data constructed by the different methods

III.6.2 Application to surgical data

The data were collected in a UK center for cardiac surgery during 1992–1998 and contain records for 6994 patients. The outcome measure is the *30-day mortality* of a patient following the operation (survival/death), and six covariates are available: *Parsonnet score*, *Age*, *Gender*, *Surgeon* (there are seven different surgeons), *Type* (there are three types of surgery operations: elective, urgent, emergency) and *Diabetes* (whether or not the patient has diabetes). The Parsonnet score, which indicates the patient's preoperative risk of death, is a well-known predictor for cardiac surgery mortality [99]. Many studies that use this dataset [100-104] fit a simple logistic regression with the

Parsonnet score as predictor. Here we will build a logistic regression tree using the Parsonnet score as predictor as well and other five covariates as splitting variables. As the true subgroups are unknown, we make a small modification to the original data to show the differences between the methods and effectiveness of the proposed method. Specifically, we first build a regular MOB tree using the original data. Treating this as the true model, we add some outliers at a node of the tree where change occurs. Then the five methods are applied to the modified data. The best method is the one that produces the true model regardless of the added outliers. The parameter settings of all the methods in tree building are the same as in the analysis of the COPD data.

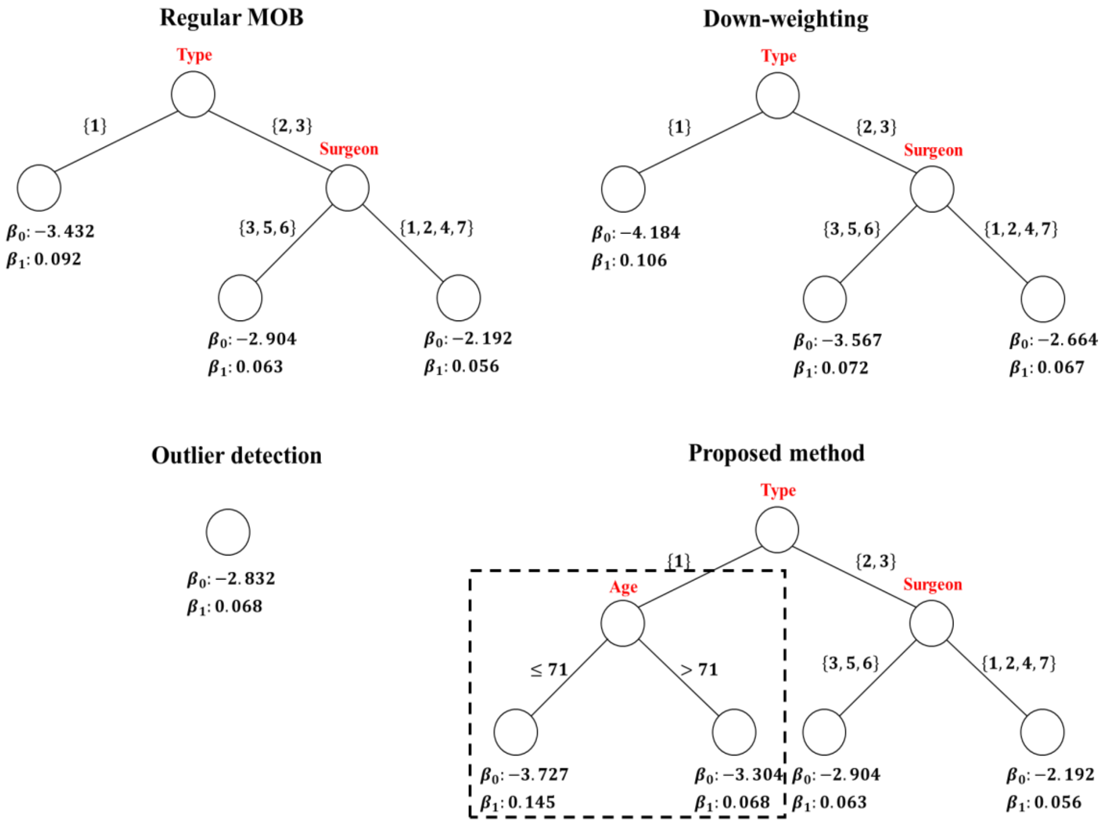


Figure III.14 The logistic regression trees for the surgical data constructed by the different methods

Figure III.14 shows the constructed trees. Outlier detection does not generate a tree. This is consistent with this method's very high miss splitting rate in Figure III.12. *Type* and *Surgeon* are split variables in all of the other trees. The proposed method also splits the node associated with *Type* = {1} by *Age*, as shown in the dashed square, while the regular MOB tree and the down-weighting tree do not. The truth regarding this node is that there is a change in the parameters of the logistic regression over *Age*, so the node should indeed be split by *Age*. However, with the added outliers, the evidence of change is masked, and thus the regular MOB algorithm and down-weighting fail to split the node, and consequently, the opportunity to capture meaningful age-related subgroups is lost. This result validates the robustness of the proposed method to outliers.

III.7 Conclusion and discussion

Logistic regression trees provide a useful method for identifying heterogeneous subgroups in binary outcome modelling. This study first uses a simulation to investigate the effects of outliers on split variable selection in building logistic regression trees. It is found that outliers slightly decrease the false splitting rate but considerably increase the miss splitting rate. A robust logistic regression tree is proposed to remedy this problem. The simulation results show that the proposed method reduces the miss splitting rate and outperforms two alternative methods in this regard, and an application to healthcare data further validates its robustness.

Another finding that deserves mention is that down-weighting and outlier detection widely used to address outliers in linear and logistic regressions, are partially useful in the regression tree context. They cannot substantially reduce and might even increase the miss splitting rate. These two methods are based on the assumption that there is a single model, so they are effective in alleviating outlier effects when the data can be explained by a single model (i.e., no splitting is needed). However, when a single model is not adequate, they may mask the evidence of change, leading to more missed splits.

Several interesting open problems related to this study remain. One problem is that outliers may also affect the cutpoint estimation after the split variable is selected. The cutpoint is found by minimizing the objective function, which involves estimating parameters for the resulting subgroups. Because outliers have an effect on parameter estimation, they may result in misleading cutpoints. We plan to explore how to treat outliers in cutpoint estimation in our future research. Another problem is how to extend the proposed robust logistic regression tree to random forests. Random forests are an ensemble method that is designed to improve the prediction performance of regular trees by generating a large number of trees and aggregating their predictions. In constructing a robust version of random forests, computational efficiency is a main concern. The method proposed in the present study involves Bayesian posterior sampling, which is computationally expensive. A recent study proposed a Bayesian logistic model using a Polya-Gamma latent variable that avoids analytic approximations and thus enhances the computational efficiency of Bayesian inference [105]. This model can also help reduce the proposed method’s computation time for the recovery of intrinsic extreme points. Incorporating this model in the proposed method will provide a potential direction regarding robust random forests.

CHAPTER IV

BINARY REGRESSION TREES FOR IMBALANCED CLASS DATA

With years of quality improvement efforts in many applications such as healthcare and manufacturing systems, the number of adverse outcomes like morality and defective product rate is gradually decreasing. It has been known that conventional binary outcome modelling methods are meant to favor the majority class, showing tendency to underestimate the probability of the minority class in prediction. To better understand such imbalanced class issue in the context of subgroup identification beyond prediction, this chapter proposes two binary regression trees for imbalanced class data. The performances of the two proposed regression tree are compared with those of logistic regression tree when outcomes of interests are (extremely) rare. This study summarizes findings from the simulation and discusses the potentials of regression trees for subgroup identification under class imbalance environment.

IV.1 Introduction

Data not contaminated by outliers do not always guarantee the satisfactory performance in binary classification. There is another practical situation of binary response data where one class has significantly fewer samples than the other class, which is called rare event data or *imbalanced class problem*. Many real-world applications like fraud detection, medical diagnosis, and healthcare informatics suffer from this problem. Usually, the cost of misclassifying the minority class (e.g., cancer) is critical compared to that of misclassifying the majority class (e.g., non-cancer) in those applications, so the prediction of the minority class is a primary concern in modelling this type of data. However, under the imbalanced class situation, most binary classification models underestimate the probability of the minority class since it favours the majority class like other binary data modelling methods. Thus, most of the research on this topic have focused on improving the detection of the minority class.

The approaches to handle imbalanced class data in literature can be roughly divided into two categories: data-level approaches and algorithm-level approaches [106]. The basic idea of data-level approaches is to make the class distribution balanced by sampling the data such that the conventional classification methods perform in the most desirable circumstance. One can randomly under-sample observations with the majority class, or artificially create synthetic samples from the minor class instead [107, 108, 109, 110]. In contrast, algorithm-level approaches adopt standard methods to treat the class imbalance. For example, different costs are assigned to different misclassification types in the objective function (e.g., cost-sensitive learning), or several models are learned from the training data and their evaluations are combined to make final prediction (e.g., ensemble methods). Sometimes, models attempt to learn the minority class samples alone (e.g., one-class classification, recognition-based methods) [111]. Neither of these two types of approaches dominates the other universally in terms of performance. In particular, cost-sensitive learning produces equivalent results to sampling methods, and it has been found that there is no difference between these two methods [112, 113].

About decision trees to handle imbalanced class data, most research have relied on sampling-based methods and cost-sensitive learning that usually jointly work with ensemble methods like random forest. Moreover, split criteria are modified or the decision boundary is adjusted to improve the predictive accuracy. For example, Drummond and Holte [114] show that the decision tree proposed by Dietterich *et al*. [115] improves the prediction power, which is originally devised to satisfy the Probably Approximately Correct (PAC) condition in tree learning. By investigating the influence of the imbalanced class on different impurity measures, Liu *et al*. [116] introduce a robust and insensitive measure to class distribution, Class Confidence Proportion (CCP), and Cieslak *et al*. [117] propose the Hellinger distance as the split criterion. Maszczyk and Duch [118] simply adopte the Renyi entropy as a split criterion and find that the Renyi entropy is effective in learning decision trees under imbalanced data with a proper choice of the order parameter α associated with the entropy. Park and Ghosh [119] extend the decision tree using the Renyi entropy under the ensemble learning framework by generating diverse trees with multiple parameters of α. Instead of using the constant α across the whole tree, Hong *et al*. [120] adaptively decide the parameter α at each node according to the class distribution. However, all of those methods have primarily prioritized the improvement on prediction accuracy alone, not on the subgroup identification.

To handle imbalanced class issue for subgroup identification, the two approaches mentioned above may not work well in the context of regression trees. As described in the previous Chapter, MOB logistic regression tree pays attention to the change in parameters of logistic regression along with covariates, and such changes are the key to identify hidden subgroups. However, the data-level approaches such as the sampling-based methods artificially reform the dataset to balance the class distribution, and thus lose the underlying structure of the original data. This may lead to misleading subgroups or fail to identify important subgroups. Regarding the algorithm-level approaches, they have to know the true class values for each observation to identify the misclassified observations and assign high cost for the observations in the learning process. If we apply the same idea for subgroup identification, we would obtain the so called misclassified subgroups. Unfortunately, subgroups are usually unknown from the beginning so that we are not able to define misclassified subgroups. In other words, the algorithm-level approaches cannot be applied for subgroup identification problem in any ways.

This chapter proposes new binary regression trees for subgroup identification while maintaining the original imbalanced class structure of the dataset. Two binary regression trees are proposed. The first tree model is a logistic regression tree assisted with Firth’s method (called Firth’s logistic regression tree hereafter) and the second one is the generalized extreme value regression tree (called GEV regression tree hereafter). This study defines suitable performance measures to assess the effect of imbalanced class on split variable selection and compares the performance of three tree models, including the MOB logistic regression tree described in the previous Chapter, Firth’s logistic regression tree, and GEV regression tree, by simulation.

The remainder of this chapter is organized as follows. Section IV.2 presents the imbalance class problem of logistic regression and related literature. Section IV.3 introduces the two proposed tree models for imbalanced binary outcomes and Section IV.4

compares performance of the MOB logistic regression tree and the two proposed trees by simulation. Finally, Section IV.5 concludes the chapter and discusses future research directions. All supplemental materials are available in Appendix B.

IV.2 Imbalance class problem in logistic regression

Logistic regression is to estimate the probability of the occurrence of binary outcomes y (e.g., survival/death) via a logistic function of K dimensional predictor $\boldsymbol{X}$. The probability $P(Y_i = 1|\boldsymbol{X}_i)$ is expressed as $\frac{exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{Ki})}{1+exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p X_{Ki})}$ with $(K + 1)$ parameters. The unknown coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_K)$ are usually obtained by maximum likelihood estimation. The maximum likelihood estimator has very desirable properties in large samples such as asymptotical unbiasedness (i.e., $E(\widehat{\boldsymbol{\beta}}_{MLE}) \approx \boldsymbol{\beta}_{true}$) and efficiency (i.e., the asymptotic variance of the maximum likelihood estimate achieves the Cramer-Rao lower bound) [121]. However, when the binary outcomes are highly imbalanced, the maximum likelihood estimates have substantial bias [122]. Specifically, King and Zeng [122] point out that the asymptotic bias on the intercept is

$$E(\hat{\beta}_0) - \beta_0 \approx \frac{\bar{\pi} - 0.5}{n\bar{\pi}(1 - \bar{\pi})}, \qquad \text{(IV.1)}$$

where $\bar{\pi}$ is the probability of the event (i.e., $Y_i = 1$) in the data. It is easy to see that $\bar{\pi}$ as well as $\bar{\pi}(1 - \bar{\pi})$ have a very small value under highly imbalanced class data and thus the bias becomes substantially large according to Equation (IV.1). In order to reduce the bias, King and Zeng subtract the $O(n^{-1})$ term of bias from the maximum likelihood estimates. In fact, Firth [123] proposes the generalized approach for removing the $O(n^{-1})$ term in the bias. Thus, Firth's approach is similar to King and Zeng's approach under imbalanced class data. Firth's idea has also been implemented to solve the separation problem (i.e., the binary outcomes are completely separated by a single predictor) in logistic regression [124]. Such bias and its effect on predictive power are covered in several studies [125-127].

Owen (2007) studies logistic regression under the infinitely imbalanced case [128]. As the number of majority class (i.e., $Y = 0$) goes infinity with a fixed number of minority class (i.e., $Y = 1$), the estimate of intercept $\hat{\beta}_0$ goes negative infinity and the estimated coefficients $\hat{\beta}_1, \dots, \hat{\beta}_K$ will approach a meaningful limit

$$\overline{X} = \frac{\int e^{X^T\beta} X dF_0(X)}{\int e^{X^T\beta} dF_0(X)}, \qquad \text{(IV.2)}$$

where $\overline{X}$ is the mean of the sample X_i corresponding to $Y = 1$ and F_0 is the distribution of X given $Y = 0$. Equation (IV.2) requires that F_0 not be a heavy-tail distribution such as Cauchy distribution and $\overline{X}$ be surrounded by F_0. These conditions are called "overlap conditions" in the paper and they are derived in the light of Silvapulle's results that characterize the existence and uniqueness of maximum likelihood estimates for logistic regression [129]. Equation (IV.2) implies that logistic regression only relies on the observations with $Y_i = 1$ via their average value of predictors in the infinitely imbalanced situation. The finding enables logistic regression to perform better by shrinking outliers toward $\overline{X}$ or clustering X_i with $Y_i = 1$ and thus fitting a logistic regression at each cluster under infinitely imbalanced class distribution.

Several logistic regression models have been developed to improve the accuracy of classification in imbalanced data. Rahayu [130] proposes AdaBoost Newton truncated regularized weighted kernel logistic regression and AdaBoost Newton truncated regularized logistic regression that show considerable improvement in the accuracy by the virtue of AdaBoost. Maalouf and Siddiqi [131] develop a rare event weighted logistic regression on large-scale imbalanced data, which performs better than the truncated regularized iteratively re-weighted least squares [132]. In addition, Maalouf and Trafalis [133] extend the weighted logistic regression assisted with a kernel method, which is suitable for small to medium sample size. Wang, Xu and Zhou [134] apply Lasso (least absolute shrinkage and selection operator) logistic regression to the unbalanced credit scoring problem. However, most of the methods focus on the fitting and predictive power of logistic regression, not logistic regression tree for subgroup identification.

IV.3 Model description

This chapter proposes two binary regression tree models for imbalanced class data. The proposed methods follow the similar learning procedure as in the previous chapter. Let Y be the binary response, $\boldsymbol{X}$ be the vector of $(K+1)$ elements (K predictors with 1 in the first element) in the binary outcome regression model, $\{Z_1, Z_2, \dots, Z_l\}$ be the set of covariates, and *n* is the sample size. Through partitioning the covariate space, the proposed regression tree explores a piecewise binary regression model $\{\mathcal{M}_b: Y \sim BM(\boldsymbol{X}; \boldsymbol{\theta_b}), b = 1, \dots, \mathcal{B}\}$ that fits observations in each subgroup b better than a global model $\{\mathcal{M}: Y \sim BM(\boldsymbol{X}; \boldsymbol{\theta})\}$, where $\boldsymbol{\theta_b}$ and $\boldsymbol{\theta}$ are model parameters of subgroup b and the population, respectively, $\mathcal{B}$ is the total number of subgroups, and "BM" stands for "binary regression model". Two binary regression models are considered: Firth' logistic regression (FL) and the generalized extreme value regression (GEV regression). For each model, two steps are involved in learning a tree: model parameter estimation and covariate space splitting. The covariate space splitting consists of two separate steps: split variable selection and cutpoint estimation. Details of the steps of each model are given as follows.

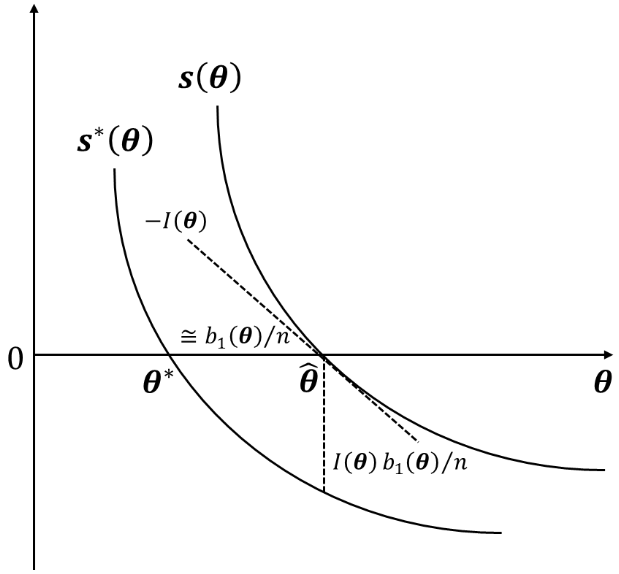
IV.3.1 Firth's logistic regression tree

IV.3.1.1. Firth's logistic regression

As explained in Chapter III, the parameters of logistic regression $\boldsymbol{\theta}$ are obtained via maximum likelihood estimation. Maximum likelihood estimates of the parameters are equivalent to the solution of the score function $\boldsymbol{s}(\boldsymbol{\theta})$, i.e., the first partial derivative of the log-likelihood $\ell(\boldsymbol{\theta};\mathcal{D})$ given data $\mathcal{D} = \{D_i = (Y_i, \boldsymbol{X}_i), i = 1, \dots, n\}$. In a regular parametric model with parameter $\boldsymbol{\theta}$, the asymptotic bias of the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ can be expanded as

$$b(\boldsymbol{\theta}) = E\left(\widehat{\boldsymbol{\theta}}\right) - \boldsymbol{\theta} = \frac{b_1(\boldsymbol{\theta})}{n} + \frac{b_2(\boldsymbol{\theta})}{n^2} + \cdots, \qquad \text{(IV.3)}$$

where n is the sample size. In order to reduce the first-order term bias, $\frac{b_1(\boldsymbol{\theta})}{n}$, in Equation (IV.3), Firth [123] proposes the modified score functions by using geometric and statistical property of the score function. The idea of the modification is illustrated in Figure IV.1. which is slightly modified from the original paper.



$\boldsymbol{s}(\boldsymbol{\theta})$
$\boldsymbol{s}^*(\boldsymbol{\theta})$
$-I(\boldsymbol{\theta})$
$\cong b_1(\boldsymbol{\theta})/n$
0
$\boldsymbol{\theta}^*$
$\widehat{\boldsymbol{\theta}}$
$\boldsymbol{\theta}$
$I(\boldsymbol{\theta})\, b_1(\boldsymbol{\theta})/n$

Figure IV.1 The idea of Firth's modification on the score function

In Figure IV.1, $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}$ are the solution of the modified score function and the original score function, respectively, and $I(\boldsymbol{\theta})$ denotes Fisher information. Firth shows that if the maximum likelihood estimates $\widehat{\boldsymbol{\theta}}$ is subject to a positive first-order bias $\frac{b_1(\boldsymbol{\theta})}{n}$, the bias can be removed by shifting the score function $\boldsymbol{s}(\boldsymbol{\theta})$ downward such that the shifted score function has a solution at $\boldsymbol{\theta}^*$. Using the property of Fisher information (i.e., the absolute value of the gradient of the score function), this idea is realized by shifting the score function $\boldsymbol{s}(\boldsymbol{\theta})$ downward by an amount $I(\boldsymbol{\theta})\ \frac{b_1(\boldsymbol{\theta})}{n}$. Thus, the modified score function is expressed as

$$\boldsymbol{s}^*(\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{\theta}) - I(\boldsymbol{\theta})\ \frac{b_1(\boldsymbol{\theta})}{n}. \qquad \text{(IV.4)}$$

In logistic regression where

$$\text{Prob}(Y_i = 1|X_i, \boldsymbol{\theta}) = \pi_i = [1 + exp\{-(\theta_0 + \textstyle\sum_{k=1}^{K} X_{ik}\theta_k)\}]^{-1}, \qquad \text{(IV.5)}$$

the first-order bias term in Equation (IV.4) takes the following form [135]:

$$b_1(\boldsymbol{\theta}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\xi}, \tag{IV.6}$$

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{W}$ is an $n \times n$ diagonal matrix whose *i*

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{W}$ is an $n \times n$ diagonal matrix whose i^{th} element is $\pi_i(1-\pi_i)$, $\boldsymbol{W\xi}$ has i^{th} element $h_i\left(\pi_i - \frac{1}{2}\right)$, and h_i is the i^{th} diagonal element of the "hat" matrix

$$\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{\frac{1}{2}}. \qquad \textbf{(}\text{IV.7)}$$

Hence, by plugging $b(\boldsymbol{\theta})$ in Equation (4.6) and $I(\boldsymbol{\theta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ into Equation (4.4), the modified score function $\boldsymbol{s}^*(\boldsymbol{\theta})$ becomes

$$\boldsymbol{s}^*(\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{\theta}) - \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\xi} \tag{IV.8}$$

Given the k^{th} component of the score function of logistic regression is $\sum_{i=1}^{n}(Y_i - \pi_i)X_{ik}$, the modified score function in logistic regression is

$$\boldsymbol{s}^*(\theta_k) = \sum_{i=1}^{n}\left\{Y_i - \pi_i + h_i\left(\frac{1}{2} - \pi_i\right)\right\}X_{ik}, \qquad (k = 0, 1, \dots, K) \tag{IV.9}$$

The estimates $\widehat{\boldsymbol{\theta}}$ are obtained via the Fisher-scoring method (which is equivalent to iteratively reweighted least squares) until parameters are converged

$$\boldsymbol{\theta}^{(\gamma+1)} = \boldsymbol{\theta}^{(\gamma)} + I^{-1}\left(\boldsymbol{\theta}^{(\gamma)}\right)\boldsymbol{s}^*\left(\boldsymbol{\theta}^{(\gamma)}\right) \tag{IV.10}$$

where γ refers to the γ^{th} iteration.

It is worth noting that Firth's idea can be understood as penalized likelihood estimation. For exponential family models such as logistic regression, the penalty term is specified by the square root of the determinant of the Fisher information evaluated at $\boldsymbol{\theta}$

(i.e., $|I(\boldsymbol{\theta})|^{\frac{1}{2}}$**),** which is known as Jeffreys' invariant prior in Bayesian framework [136]. The penalty term approaches zero as the sample size goes to infinity, while it removes the $\mathrm{O}(n^{-1})$ term in bias in the maximum likelihood estimation for small and imbalanced samples. The penalized log-likelihood for logistic regression is represented as

$$\ell(\boldsymbol{\theta};\mathcal{D})^{*} = \ell(\boldsymbol{\theta};\mathcal{D}) + \frac{1}{2}ln|I(\boldsymbol{\theta})|. \qquad \text{(IV.11)}$$

Then, the score function of the penalized likelihood is expressed as

$$\boldsymbol{s}^{*}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\left(\ell(\boldsymbol{\theta};\mathcal{D}) + \frac{1}{2}ln|I(\boldsymbol{\theta})|\right) = \boldsymbol{s}(\boldsymbol{\theta}) + \frac{1}{2}trace\left[I(\boldsymbol{\theta})^{-1}\left\{\frac{\partial I(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\}\right], \qquad \text{(IV.12)}$$

which is equivalent to Equation (4.9) by simple algebra. It is known that Firth's penalized maximum likelihood estimation shrinks the maximum likelihood estimates toward zero [137], and thus reduces both bias and variance in the estimation of imbalanced data [138]. Unfortunately, the direct use of Firth's estimates is not possible in the parameter instability test for split variable selection, since Firth's approach produces the penalized likelihood estimates, not maximum likelihood estimates. This implies that further modification is required to apply this approach in the parameter instability test, which will be described in the next Section.

Once parameters are estimated at each node, the Firth's logistic regression tree is constructed by following the same procedure of the MOB tree described in Chapter III: split variable selection and cutpoint estimation for subgroup identification. In the next Section, the procedure are briefly covered again in the context of Firth's logistic regression tree.

IV.3.1.2. Covariate space splitting of Firth's logistic regression tree

Split variable is selected by parameter instability test. At the current node R, the test is conducted for each covariate $Z_j, j = 1, \dots, l$, to determine whether the node R is split or not. If a global Firth's logistic regression model with parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ fits the data well, the node R is not split by Z_j. The parameter is said to be stable with respect to Z_j in this case. If a global Firth's logistic regression model is inadequate to fit the data, the node R should be better split into two subgroups

$$R_1 = \{\boldsymbol{X}|Z_j \leq c\}, R_2 = \{\boldsymbol{X}|Z_j > c\},$$

by the cutpoint of c of Z_j. Thus, two separate Firth's logistic regression models are fitted into the subgroups with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. In this case, the parameter at the node R is said to be unstable with respect to Z_j. Naturally, the covariate with the most convincing evidence for unstable parameter is selected as split variable.

The tree splitting at each node R is formulated as the following hypothesis testing problem

$$\begin{aligned} H_0 &: Y_i \sim FL(\boldsymbol{X}_i; \boldsymbol{\theta}_0) \quad for\ i = 1, \dots, n \\ H_1 &: Y_i \sim \begin{cases} FL(\boldsymbol{X}_i; \boldsymbol{\theta_1}) & if\ \boldsymbol{X}_i \in R_1 \\ FL(\boldsymbol{X}_i; \boldsymbol{\theta_2}) & if\ \boldsymbol{X}_i \in R_2 \end{cases} \end{aligned} \qquad \text{(IV.13)}$$

where n is the sample size of available data at node R. Essentially, this hypothesis testing is equivalent to the problem of change detection in parameter over covariate $\{Z_1, Z_2, \dots, Z_l\}$. In the test, the smaller p-value is, the more plausible the node R should be split. Thus, the covariate with the smallest p-value is selected as the split variable at node R. As we know, the parameter instability in Equation (IV.13) is assessed by a score-based test, which belongs to the class of generalized M-fluctuation tests. At node R, the parameters of the global Firth's logistic regression model are estimated by minimizing the negative penalized log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D})^*$ in Equation (IV.11), resulting in the penalized maximum likelihood estimate

$$\widehat{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \textstyle\sum_{i=1}^{n} \ell(\boldsymbol{\theta}; D_i)^*, \tag{IV.14}$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_K)$ and $D_i = (Y_i, \boldsymbol{X}_i), i = 1, \dots, n$ are available data at node R. This is equivalent to the solution of the modified score function in Equation (IV.8)

$$\textstyle\sum_{i=1}^{n} \boldsymbol{s}^*\left(\widehat{\boldsymbol{\theta}}^*; D_i\right) = \boldsymbol{0}. \tag{IV.15}$$

However, the modified score does not have an expectation of zero at the true parameter value. This means that Equation (IV.15) cannot be directly used for the parameter instability test. In other words, the empirical cumulative score process $\boldsymbol{W}_j(t)$ based on the modified score function in Equation (IV.15) does not converge to a Brownian bridge $\boldsymbol{W}^0$ under the null hypothesis. To address this issue, this study uses the penalized maximum likelihood estimates $\widehat{\boldsymbol{\theta}}^*$ under the original score function $\boldsymbol{s}\left(\widehat{\boldsymbol{\theta}}^*; D_i\right)$. Note that $\widehat{\boldsymbol{\theta}}^*$ is not the solution of the original score function, so the sum of scores evaluated at $\widehat{\boldsymbol{\theta}}^*$ is no longer zero on all the observations, i.e., $\sum_{i=1}^{n} \boldsymbol{s}\left(\widehat{\boldsymbol{\theta}}^*; D_i\right) \neq \boldsymbol{0}$. This study applied a simple modification that subtracts the mean score from the individual scores, to make the resulting scores sum to zero. The corrected score function is

$$\textstyle\sum_{i=1}^{n} \boldsymbol{s}^c\left(\widehat{\boldsymbol{\theta}}^*; D_i\right) = \sum_{i=1}^{n} \left\{ \boldsymbol{s}\left(\widehat{\boldsymbol{\theta}}^*; D_i\right) - \frac{\sum_{i=1}^{n} \boldsymbol{s}\left(\widehat{\boldsymbol{\theta}}^*; D_i\right)}{n} \right\} = \boldsymbol{0} \tag{IV.16}$$

By leveraging the corrected score function in Equation (IV.16), the empirical cumulative score process of Firth's logistic regression tree along with Z_j is established to detect systematic patterns of scores

$$\boldsymbol{W}_j^{Firth}(t) = \hat{\boldsymbol{J}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{s}^{\boldsymbol{c}}\left(\hat{\boldsymbol{\theta}}^*; D_{(i|Z_j)}\right),\ 0 \le t \le 1 \qquad \text{(IV.17)}$$

where $\hat{\boldsymbol{J}}$ is an estimate of the covariance matrix of the corrected scores, $\lfloor nt \rfloor$ is the integer part of nt, and $D_{(i|z_j)}$ is the observation with the ith smallest value of Z_j. As explained in the Chapter III, $\sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{s}^{\boldsymbol{c}}\left(\hat{\boldsymbol{\theta}}^*; D_{(i|Z_j)}\right)$ reflects the overall lack of fit up to the *nt*th observation in Firth logistic regression at node R. A suitable estimate of the covariance matrix $\hat{J}$ can be outer product of gradient (OPG) or the observed information matrix. The inverse square root of the covariance matrix, i.e., $\hat{\boldsymbol{J}}^{-1/2}$, in Equation (IV.17) decorrelates the scores of the $(K + 1)$ parameters, so that we can inspect the score of individual parameter separately. Thus, $\boldsymbol{W}_j^{Firth}(t)$ in Equation (IV.17) captures deviations from the null hypothesis (e.g., H_0 in Equation (IV.13)) of parameter stability.

The test statistic $\mathcal{T}$ for the test in Equation (IV.13) can be derived in the same fashion as the MOB logistic regression tree by applying some scalar functional λ to $\boldsymbol{W}_j^{Firth}(t)$. Depending on the nature of Z_j, two specific forms of $\mathcal{T}$ are given below:

$$\mathcal{T} = \lambda\left(\boldsymbol{W}_j^{Firth}(t)\right) = \max_{i=L,\dots,U} \left(\frac{i}{n} \cdot \frac{n-i}{n}\right)^{-1} \left\|\boldsymbol{W}_j^{Firth}\left(\frac{i}{n}\right)\right\|_2^2 \text{ for a continuous } Z_j, \qquad \text{(IV.18)}$$

$$\mathcal{T} = \lambda\left(\boldsymbol{W}_j^{Firth}(t)\right) = \sum_{q=1}^{Q} \frac{|I_q|}{n}^{-1} \left\|\boldsymbol{W}_j^{Firth}\left(\frac{i}{n}\right)\right\|_2^2 \text{ for a categorical } Z_j. \qquad \text{(IV.19)}$$

The meaning of notations in both Equations (IV.18) and (IV.19) are the same as that of the MOB in Chapter III. The limiting distribution of the test statistics in Equations (IV.18) and (IV.19) is the supremum of a squared, $(K+1)$-dimensional tied-down Bessel process, and χ^2 distribution with $(K+1)(Q-1)$ degrees of freedom where Q is the number of classes in a categorical covariate Z_j, respectively. Using appropriate test statistic $\mathcal{T}$, the parameter instability test is performed for each Z_j and the covariate with the minimal p-value less than a pre-determined significance level α corrected for multiple (totally l) testings is selected as the split variable.

After the split variable is determined, the optimal cutpoint c is computed to form two subgroups. This cutpoint will lead to the most heterogeneous parameter values between subgroups as much as possible. Over all conceivable cutpoints of the split variable, the optimal point can be obtained by locally optimizing the negative log-likelihood, yielding maximal reduction in the negative log-likelihood before and after split.

IV.3.2 Generalized extreme value regression tree

IV.3.2.1. Generalized extreme value (GEV) regression

A logistic regression model uses a symmetric link function whose rate of approaching each class is identical for modelling the response curve of probability. Such identical rate may not perform well in estimating the probability of the minority class. In order to overcome the issues, Calabrese and Osmetti (2013) [139] propose the generalized extreme value (GEV) regression model for imbalanced outcome. This model adopts the quantile function (i.e., the inverse of cumulative distribution function) of the GEV distribution to model the probability response curve. It is known that the GEV distribution is very flexible with a shape parameter τ that controls the shape and the size of the tails of distribution, which essentially leads to an asymmetric link function that is able to handle imbalanced class data in a flexible way. The cumulative distribution function of the GEV distribution is given by

$$F_X(x) = \exp\left\{-\left[1+\tau\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\tau}\right\},\ -\infty < \tau < \infty,\ -\infty < \mu < \infty,\ \sigma > 0 \qquad \text{(IV.20)}$$

where τ is a shape parameter, while μ and σ are location and scale parameters, respectively. Depending on the sign and value of τ, special cases can be recovered: Gumbel distribution ($\tau \to 0$), Frechet distribution ($\tau > 0$), and Weibull distribution ($\tau < 0$). In particular, the cumulative Gumbel distribution is the log-log function in binary response modelling. Then, the probability $\pi(\boldsymbol{X}_i) = P(Y_i = 1|\boldsymbol{X}_i)$ in GEV regression model is defined as

$$P(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta}) = \pi(\boldsymbol{X}_i) = \exp\{-[1+\tau(\boldsymbol{X}_i^T\boldsymbol{\beta})]^{-1/\tau}\},\ i = 1,2, \dots., n \qquad \text{(IV.21)}$$

and the link function of the model is given by

$$\frac{\{-\ln[\pi(\boldsymbol{X}_i)]\}^{-\tau}-1}{\tau} = \boldsymbol{X}_i^T\boldsymbol{\beta},\ i = 1,2, \dots., n \qquad \text{(IV.22)}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients (i.e., $\boldsymbol{\beta} = \left[\beta_0, \beta_1, \dots, \beta_p\right]^T$**)** .

For parameter estimation, maximum likelihood estimation is used. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau)$, then the log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left\{-Y_i[1+\tau(\boldsymbol{X}_i^T\boldsymbol{\beta})]^{-\frac{1}{\tau}} + (1-Y_i)\ln\left[1-\exp\left[-[1+\tau(\boldsymbol{X}_i^T\boldsymbol{\beta})]^{-\frac{1}{\tau}}\right]\right]\right\}. \qquad \text{(IV.23)}$$

Equation (IV.23) exists only for $\{\boldsymbol{X}_i: 1 + \tau(\boldsymbol{X}_i^T\boldsymbol{\beta}) > 0\}$, and it is maximized by optimization algorithms with the specification of initial values. Since the Fisher information is not a diagonal matrix (i.e., the parameters $\boldsymbol{\beta}$ and τ are dependent), $\boldsymbol{\beta}$ and τ have to be estimated simultaneously. The original paper suggests the following initial values for optimization: $\tau^* \cong 0$, $\beta_k^* = 0$ for $k = 1, \dots, p$ and $\beta_0^* = \ln[-\ln(\bar{y})]$. The Fisher information is given in Appendix B.1.

The beauty of the GEV regression in the context of tree model is that it enables flexible modeling over different class ratios at different subgroups. In other words, GEV regression tree offers freedom for the choice of links according to the observations at each subgroup, so it accommodates the different degrees of imbalanced class. This flexibility lays the very foundation for constructing subgroup-specific models.

After estimating the parameters of GEV regression, the remaining procedure for covariate space splitting is identical with the previous proposed tree model. Thus, we skip the general procedure of tree learning, but focus on the score functions of GEV regression tree and degree of the limiting distribution instead in the next Section.

IV.3.2.2. Covariate space splitting of GEV regression tree

The tree splitting at each node R is formulated as the following hypothesis testing problem

$$\begin{aligned} H_0 &: Y_i \sim GEV(\boldsymbol{X}_i; \boldsymbol{\theta}_0) \quad for\ i = 1, \dots, n \\ H_1 &: Y_i \sim \begin{cases} GEV(\boldsymbol{X}_i; \boldsymbol{\theta}_1) & if\ \boldsymbol{X}_i \in R_1 \\ GEV(\boldsymbol{X}_i; \boldsymbol{\theta}_2) & if\ \boldsymbol{X}_i \in R_2 \end{cases} \end{aligned} \tag{IV.24}$$

where n is the sample size of available data at node R and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau) = \left(\beta_0, \beta_1, \dots, \beta_p, \tau\right)$. For the parameter instability test in GEV regression, we should consider the score function of the shape parameter τ as well as that of parameters associated with predictors. As we know, the score function is defined as the partial derivatives of the (negative) log-likelihood

$$\boldsymbol{s}(\boldsymbol{\theta}; D_i) = \frac{\partial \ell(\boldsymbol{\theta}; D_i)}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ell(\boldsymbol{\theta}; D_i)}{\partial \boldsymbol{\beta}}, \dots, \frac{\partial \ell(\boldsymbol{\theta}; D_i)}{\partial \tau}\right). \tag{IV.25}$$

Specifically, the score functions of regression coefficient β_j and shape parameter τ in GEV regression are given by

$$\frac{\partial l(\boldsymbol{\beta}, \tau)}{\partial \beta_k} = -\sum_{i=1}^{n} X_{ik} \frac{\ln[\pi(\boldsymbol{X}_i)]}{1+\tau\left(\boldsymbol{X}_i^T \boldsymbol{\beta}\right)} \frac{Y_i - \pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)}, \qquad k = 0,1, \dots, K, \tag{IV.26}$$

$$\frac{\partial l(\boldsymbol{\beta}, \tau)}{\partial \tau} = \sum_{i=1}^{n} \left[\frac{1}{\tau^2} \ln(1 + \tau \boldsymbol{X}_i^T \boldsymbol{\beta}) - \frac{\boldsymbol{X}_i^T \boldsymbol{\beta}}{\tau\left(1+\tau \boldsymbol{X}_i^T \boldsymbol{\beta}\right)}\right] \frac{y_i - \pi(\boldsymbol{X}_i)}{1-\pi(\boldsymbol{X}_i)} \ln[\pi(\boldsymbol{X}_i)], \tag{IV.27}$$

respectively, and $\widehat{\boldsymbol{\theta}}$ is the solution of the $(K+2)$ score equations $\sum_{i=1}^{n} \boldsymbol{s}\left(\widehat{\boldsymbol{\theta}}; D_i\right) = \boldsymbol{0}$ associated with the K predictors, one intercept, and the shape parameter τ.

Using the score function of GEV regression, the empirical cumulative score process along with Z_j is expressed as

$$\boldsymbol{W}_j^{GEV}(t) = \widehat{\boldsymbol{J}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor nt \rfloor} \boldsymbol{s}\left(\widehat{\boldsymbol{\theta}}; D_{(i|Z_j)}\right), \ 0 \leq t \leq 1 \tag{IV.28}$$

where $\widehat{\boldsymbol{J}}$ is an estimate of the covariance matrix of the scores, $\lfloor nt \rfloor$ is the integer part of nt, and $D_{(i|z_j)}$ is the observation with the ith smallest value of Z_j. Here, the inverse square root of the covariance matrix, $\widehat{\boldsymbol{J}}^{-1/2}$, in Equation (IV.28) decorrelates the scores of the $(K+2)$ parameters so that we can inspect the score of individual shape parameter as well as parameters associated with predictors separately. Likewise, $\boldsymbol{W}_j^{GEV}(t)$ in Equation (IV.28) captures deviations from the null hypothesis of parameter stability.

The test statistic for the test in Equation (IV.24) is the same as that of the test in Firth’s logistic regression tree, which is given as follows:

$$\mathcal{T} = \lambda\left(\boldsymbol{W}_j^{GEV}(t)\right) = \max_{i=L,\dots,U} \left(\frac{i}{n}\cdot\frac{n-i}{n}\right)^{-1} \left\|\boldsymbol{W}_j^{GEV}\left(\frac{i}{n}\right)\right\|_2^2 \text{ for a continuous } Z_j, \qquad \text{(IV.29)}$$

$$\mathcal{T} = \lambda\left(\boldsymbol{W}_j^{GEV}(t)\right) = \sum_{q=1}^{Q} \frac{|I_q|}{n}^{-1} \left\|\boldsymbol{W}_j^{GEV}\left(\frac{i}{n}\right)\right\|_2^2 \text{ for a categorical } Z_j \qquad \text{(IV.30)}$$

The only difference lies in the number of degrees in the limiting distribution of the test statistic, which is associated with the additional shape parameter. For a continuous covariate, the limiting distribution of test statistic is the supremum of a squared, $(K+2)$-dimensional tied-down Bessel process. For a categorical covariate, the limiting distribution of test statistic is χ^2 distribution with $(K+2)(Q-1)$ degrees of freedom. Using appropriate test statistic $\mathcal{T}$ and limiting distribution, the parameter instability test is performed for each Z_j and the covariate with the minimal p-value less than a pre-determined significance level α corrected for multiple (totally l) testings is selected for a split. The remaining step for finding the optimal cutpoint c is exactly the same as that of the MOB and Firth’s logistic regression tree.

IV.4 Performance comparison

IV.4.1 Setup for the simulation

The simulation setup for this study is similar as that of Chapter III. In this Chapter, we consider a scenario where there are three predictors $\boldsymbol{X} = (X_1, X_2, X_3)$ and a single time covariate T. Unlike the setting of the previous Chapter where a single predictor is used in simulation, three predictors are considered here. This is to generate a little more complicated predictor spaces, which can avoid the complete separation problem in binary regression. Two scenarios are simulated, as illustrated in Figure IV.2: *no-change case* and the *change case*. For each scenario, a dataset is first generated from the logistic regression model $log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ with ten different degrees of class balance in binary outcomes. Then, the parameter instability test with the statistic in Equations (III.6), (IV.18) and (IV.29) is conducted for three tree models, including the MOB logistic regression tree, Firth's logistic regression tree, and GEV regression tree, to decide whether to split the data by the covariate T. Consequently, the effects of degree of class balance are investigated by assessing the performance of the parameter instability test in finding the true underlying scenario for the data.

Specifically, X_1 follows a normal distribution with a mean of 0.7 and a standard deviation of 0.7, X_2 follows a normal distribution with a mean of -0.5 and a unit standard deviation, and X_3 follows a continuous uniform distribution over $[0.5,\ 1.2]$. In the no-change case, for the parameter $\boldsymbol{\theta}_0 = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_3)$, β_0 is used to control the degrees of balance, while $\beta_1,\ \beta_2$ and β_3 take fixed values -1, 0.5 and 0.5, respectively. Ten different degrees of balance, $\{1\%,\ 5\%,\ 10\%,\ 15\%,\ 20\%,\ 25\%,\ 30\%,\ 35\%,\ 40\%,$

$45\%\}$, are considered, which is realized by the ten different values of parameter $\beta_0 = \{-4.02,\ -2.59,\ -1.85,\ -1.32,\ -0.98,\ -0.72,\ -0.36,\ -0.17,\ 0.06,\ 0.33\}$, on average. In the change case, the pre-change parameter $\boldsymbol{\theta}_1$ and post-change parameter $\boldsymbol{\theta}_2$ take the same values in β_1, β_2 and β_3 as the parameter $\boldsymbol{\theta}_0$, while β_0 takes different values depending on the degrees of balance and change. Six different degrees of change, $\{5\%,\ 10\%,\ 20\%,\ 40\%,\ 60\%,\ 80\%\}$ in β_0, are considered, and the performance is compared at each degree of balance. The parameter values of β_0 in $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ over different degrees of balance are given in Appendix B.2.

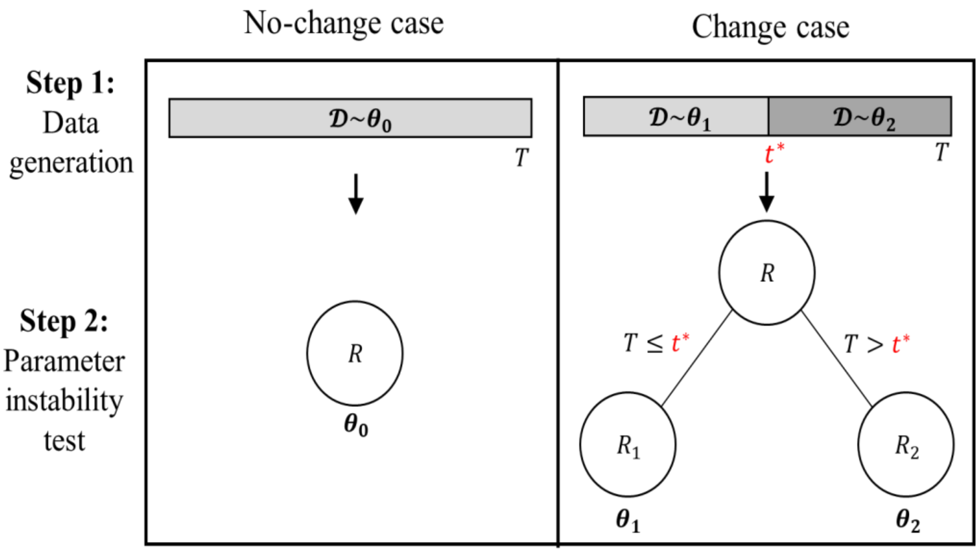


Figure IV.2 The two scenarios considered in the simulation study and the corresponding node splitting in the parameter instability test over different degrees of balance

The same performance measures in Chapter III are used for the two scenarios: the probability of a Type I error or the false splitting rate for no-change case and the probability of a Type II error or the miss splitting rate for the change case. To assess the performance in each scenario, 1000 runs were carried out under the significance level of

0.05, and the percentage of runs with false splitting or miss splitting was calculated. The results are summarized in the following subsections.

IV.4.2 Effect of class imbalance in the no-change case

Figure IV.3 shows the results for the no-change case. In Figure IV.3, the Y-axis represents the false splitting rate, and the X-axis represents the degree of balance. Four parameter instability tests are compared in the simulation: 1) test of MOB logistic regression tree with respect to β_0 (i.e., LRT with respect to β_0), 2) test of Firth's logistic regression tree with respect to β_0 (i.e., FLRT with respect to β_0), 3) test of GEV regression tree with respect to β_0 (i.e., GEVR with respect to β_0) and 4) test of GEV regression tree with respect to τ and β_0 (i.e., GEVR with respect to τ and β_0). Since the degree of balance is controlled by the intercept β_0 in data generation, parameter instability tests should be carried out over β_0 in each model. However, the shape parameter τ and the intercept β_0 (as well as parameters associated with predictors) are correlated in GEV regression as seen in Equation (IV.21). Thus, it is worthwhile to consider both the shape parameter τ and the intercept β_0 in parameter instability tests.

When the class in the data is highly imbalanced (i.e., low degree of balance), the false splitting rate of both MOB and Firth's logistic regression tree are around 0.05, which is consistent with the specified significance level of 0.05 in the parameter instability test. However, parameter instability tests of GEV regression tree tend to have higher false splitting rate in this case. In particular, the false splitting rates of GEVR with respect to τ and β_0 are uniformly greater than those of MOB and Firth's logistic regression tree, and they hit the highest point at the degree of balance of 1%. The false splitting rate of GEVR with respect to β_0 alone (i.e., red line) is around 0.05 when the degree of balance is equal to and greater than 5%, but it soars at the degree of balance of 1%. These indicate that the false splitting rate of the GEV regression tree is higher than the specified significant level in presence of highly imbalanced class. The next subsection provides intuitive explanations for this.

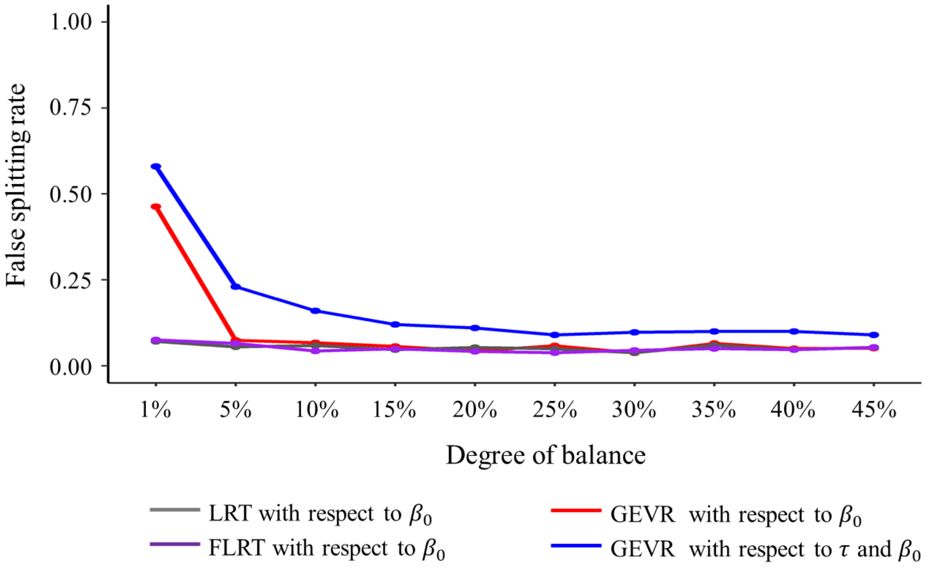


Figure IV.3 Results for the false splitting rate of the four parameter instability tests in the no-change case

IV.4.3 Evidence for high false splitting rate of GEV regression tree

To provide an intuitive understanding of the high false splitting rate of the GEV regression tree in cases of highly imbalanced class, Figure IV.4 gives a simulated example of scores in the parameter instability test in the no-change case, where observations with $Y = 1$ and $Y = 0$ are denoted by triangles and circles, respectively, and the scores from parameter instability tests with respect to τ are marked in red and the scores from the tests with respect to β_0 are marked in blue, respectively. In Figure IV.4, both scores still randomly fluctuate around zero, but the scores of τ exhibits larger variation compared to that of β_0, leading to stronger evidence of parameter instability. As a result, the false splitting rate over the shape parameter τ is higher than that over the intercept β_0.

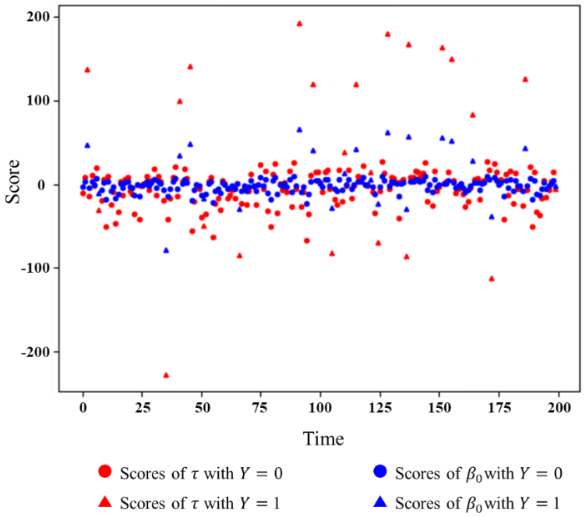


Figure IV.4 A simulated example to illustrate the high splitting rate of $\boldsymbol{\tau}$

Such high false splitting rate of the shape parameter τ is caused by the large variance of the estimated $\hat{\tau}$ from maximum likelihood estimation. Figure IV.5 shows the collection of the estimated $\hat{\beta}_0$ and $\hat{\tau}$ whose values achieve the maximum likelihood or very close values to the maximum likelihood of the simulated data. From the optimization perspective, this high variability implies the existence of multiple optima. In the particular example, the likelihood function is highly flat around the maximum likelihood estimates, which can be captured by Fisher information evaluated at the maximum likelihood estimates. In other words, there are many other good estimates compared to the resulting optimum. In this case, the parameter instability test may falsely consider other candidates better than the current estimates and thus leads to large false splitting rate.

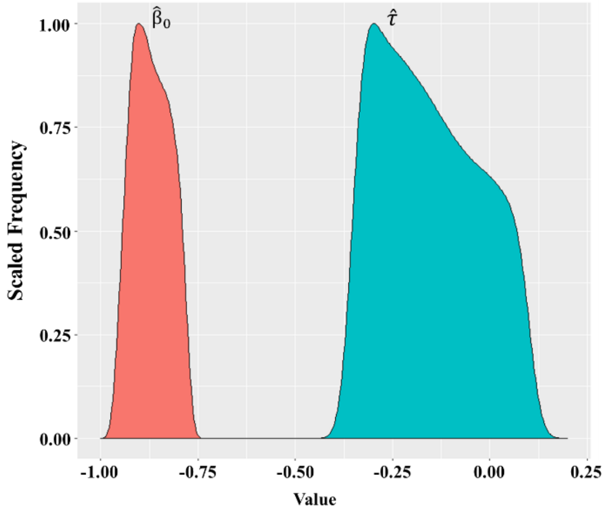


Figure IV.5 Empirical distribution of the estimates $\boldsymbol{\hat{\beta}_0}$ and $\boldsymbol{\hat{\tau}}$

In fact, the large variance of the maximum likelihood estimator in the generalized extreme value distribution has been discussed in the literature. When the shape parameter τ is not zero, the support of the distribution depends on the parameters [140]. In this case, the maximum likelihood estimators are applicable, but tend to lose their asymptotic properties, which leads to very large variance of the estimates [141]. Based on the explanation above, such large variance in estimators results in the false splitting rate greater than 0.05 across all the degrees of balance. To be specific, when the absolute value of the shape parameter is greater than 0.5, the maximum likelihood estimators do not satisfy regularity conditions (i.e., non-regular condition) and do not always exist [142, 143]. In our simulation, most of the absolute values of the estimated shape parameter in simulation are greater than 0.5 when data is highly imbalanced in class (i.e., degree of balance = 1%). Due to non-regular situation of the maximum likelihood estimates, the variance of the estimated shape parameter $\hat{\tau}$ becomes so large, resulting in substantially high false splitting rate. Since the shape parameter and the intercept term are correlated in

GEV regression, the non-regular behavior of the shape parameter partially influences on the intercept as well. Thus, GEV regression with respect to β_0 shows higher false splitting rate in highly imbalanced cases too. In summary, the high false splitting rate of GEV regression trees is essentially due to the fact that the estimation of the shape parameter τ is unstable by nature. In other words, the shape parameter tends to be sensitive to the data and over-accommodate the degree of balance.

IV.4.4 Effect of class imbalance in the change case

Figure IV.6 shows the results for the change case. The six panels represent different magnitudes of change. In each panel, the miss splitting rates of LRT and FLRT are comparable over different degrees of balance. For smaller changes (i.e., degree of change = 5%, 10% and 20%), all models show high miss splitting rates. This is simply because it is hard to detect a small change by nature. Compared to LRT and FLRT, GEV regression tree (i.e., blue and red lines) show uniformly lower miss splitting rate across the entire degrees of balance. This means that GEV regression tree performs better in detecting lower changes compared to LRT and FLRT regardless of the degree of class balance in the data. Note that the lower miss splitting rates of GEV regression tree at the degree of balance of 1% are not due to its high detection power but due to the effect of undesirable maximum likelihood estimates (i.e., non-regular problem) mentioned in subsection IV.4.3. For larger changes (i.e., degree of change = 40%, 60% and 80%), when the proportion of the minority class is equal to and greater than 10%, the miss splitting rate of GEV regression with respect to β_0 is close to or slightly higher than that of LRT and FLRT. Given that the false splitting rate of GEV regression tree over the intercept β_0

is around 0.05 at these degrees of balance as shown in Figure IV.3, the GEV regression tree is recommended for identifying subgroups, especially when the proportion of the minority class is greater than 5%.

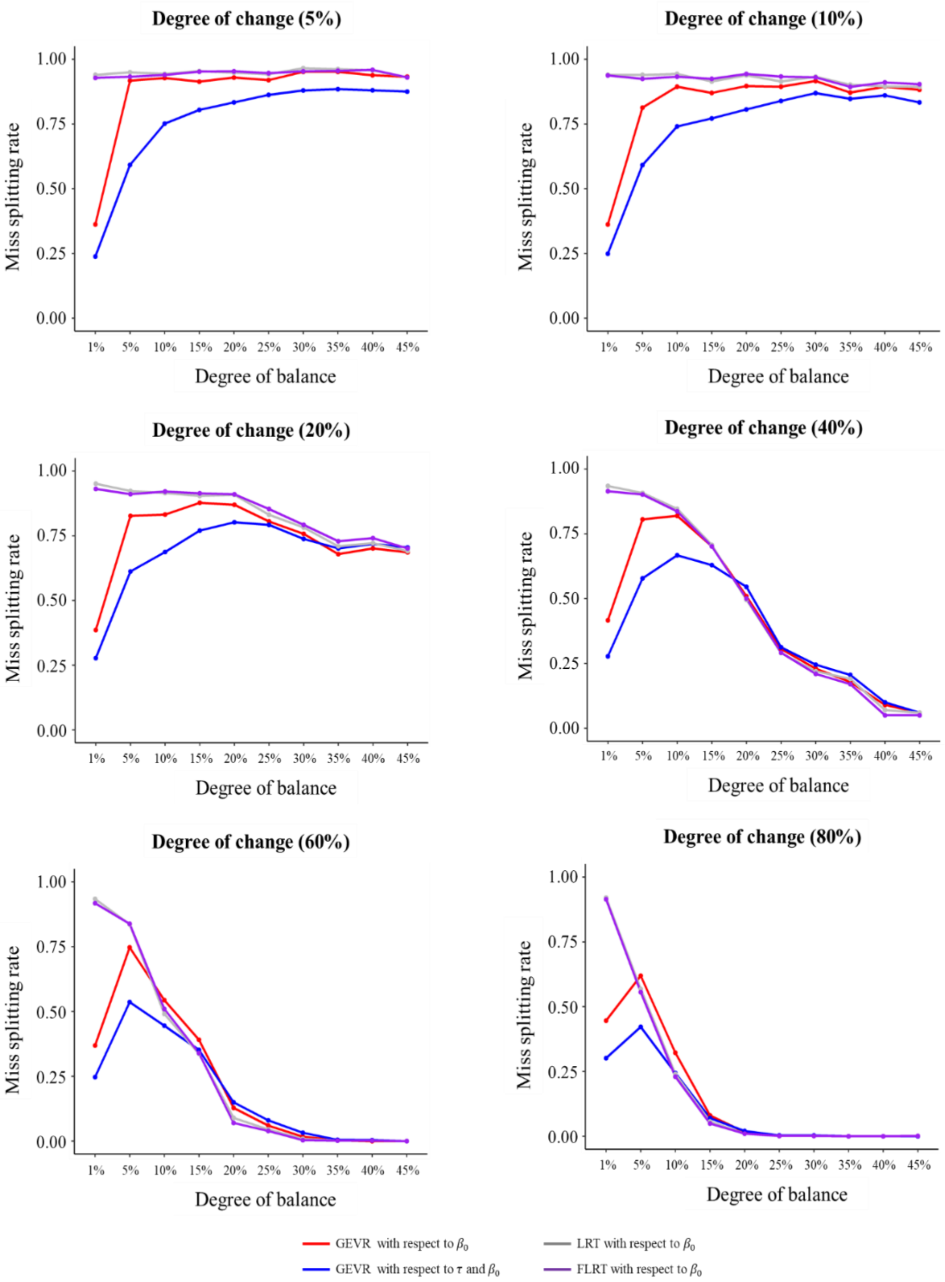


Figure IV.6 Results for the miss splitting rate of the four parameter instability tests in the change case

IV.4.5 Summary of findings from the simulation

This section summarizes the findings from the simulation and provides guidelines on the use of GEV regression tree in practice.

1) Parameter instability tests of Logistic and Firth's logistic regression for split variable selection are not influenced by the degree of balance under the no-change case. However, the tests of GEV regression tree with respect to τ and β_0 produces higher false splitting rate across all degrees of balance. This is due to the large variance of the τ estimator in maximum likelihood estimation.

2) When classes are extremely imbalanced (i.e., degree of balance = 1%), the absolute value of the estimated shape parameter $\hat{\tau}$ is greater than 0.5 in most cases. This situation makes the maximum likelihood estimator lose the desirable asymptotic properties (i.e., non-regular situation) and results in higher false splitting rate than otherwise.

3) For miss splitting rate, MOB logistic regression tree and Firth's logistic regression tree are comparable to each other regardless of the degrees of balance in class and degrees of change.

4) GEV regression tree conducting parameter instability test over the intercept produces slightly lower miss splitting rate than those of MOB and Firth's logistic regression trees when data is quite imbalanced and degree of change is quite low. For large changes, the three regression trees are comparable in terms of miss splitting rate.

5) With regard to the shape parameter, GEV regression tree should be used when the maximum likelihood estimate of the shape parameter is between –0.5 and 0.5. In fact, it is known that the shape parameter usually lies in the range between –0.5 and 0.5 in practice [144, 145, 146] and we should follow the same guidance when constructing GEV regression tree. Over the plausible range, GEV regression tree evaluates the legitimate heterogeneous effects of predictors on the probability of event in the imbalanced class situation and thus achieves flexible modelling over different class ratios at different subgroups.

IV.5 Concluding remarks

Our study proposes two binary regression trees for subgroup identification under imbalanced class data. We use a simulation to investigate the effects of the degree of balance in class on the performance of three regression trees: logistic regression tree, Firth's logistic regression tree and generalized extreme value regression tree. It is found that false splitting rates of MOB and Firth's logistic regression tree are not influenced by imbalanced class distribution. However, GEV regression tree shows high false splitting rate when classes are extremely imbalanced. For miss splitting rate, MOB logistic regression tree and Firth's logistic regression tree are comparable across all degrees of balance, while GEV regression tree makes small improvement when data is quite imbalanced and degree of change is quite low. For large magnitudes of change, all three regression trees are comparable except for the GEV regression tree with the parameter instability test over the shape parameter and intercept. Through this simulation, we recommend that the GEV regression tree should be constructed in the general case of class imbalanced data, and changes in the coefficients instead of the shape parameter should be considered when the GEV tree is used.

There are several interesting future research directions on this topic. First, Wang and Dey [147] also propose the generalized extreme value distribution as a link function. Instead of using maximum likelihood estimation, they estimate parameters via the Bayesian framework using normal distributions as priors of model parameters. Since their approach does not rely on maximum likelihood estimation on parameters, they obtain viable estimate of the shape parameter in the situation where maximum likelihood estimation breaks down and thus incorporate a wide range of skewness and flexibility in modeling the binary response curve. With a proper method for split variable selection, Bayesian GEV regression tree can be potentially used for subgroup identification under imbalanced class data. Second, Agarwal *et al*. [148] find that the GEV link with log loss results in a non-convex optimization problem, so they propose a GEV link with canonical loss to guarantee convexity for any value of the shape parameter. Zhang *et al*. [149] also propose a GEV link with convex loss to handle imbalanced class issues. Those two models can be alternatives to the GEV regression used in this study and extended into the tree framework with residuals-based hypothesis testing for selecting split variables like GUIDE [28].

CHAPTER V

CONCLUSIONS

V.1 Summary of contributions

The overarching goal of my scholarly work is to develop subgroup identification models based on the integration of statistical modelling and machine learning techniques. The proposed models in this dissertation can ferret out more informative data from the hidden subgroups for the population of interest. Specifically, this dissertation has been focused on handling special aspects of practical problems for subgroup identification. The main contributions are summarized as follows.

V.1.1 Correlation tree for subgroup identification

Correlation is the measure to quantify the strength of the relationship between two variables. It is natural that correlation depends on the condition of other covariates, so it is imperative to identify subgroups with different correlation measures in the population. However, the subgroup is discovered by manual specification in current practices, which is not efficient and may miss potential covariates whose effects are unknown. In Chapter II, we develop a correlation tree for automatic subgroup identification and provides meaningful objective functions to meet various needs in practice. The effectiveness of the correlation tree is demonstrated by the case study in neural correlate studies, but the proposed model is broadly applicable to other fields where correlation is a main concern.

V.1.2 Robust logistic regression tree for subgroup identification in healthcare outcome modeling

Collecting outcome data is a routine in healthcare practices to assess and improve the quality of care providers. Outcome measures usually have different relationships with other covariates such as physiological and treatment variables of patients. Thus, subgroup-wise outcome modelling is indispensable and logistic regression trees serve as this purpose. However, real-world data are often contaminated by aberrant observations such as outliers, and most studies have addressed outlier problems with respect to model fitting, not subgroup identification. In Chapter IV, this dissertation thoroughly investigates the outlier problem in the context of discovering subgroups by logistic regression trees. The contribution of this study is to reveal the effect of outliers on subgroup learning and develop robust logistic regression tree for identifying subgroups. By comparing the performance of the proposed method with two methods that extend the conventional ideas for addressing outliers in logistic regression to the tree context, this research provides deep understanding of the conventional ideas and demonstrates the effectiveness of the proposed method.

V.1.3 Binary regression trees for imbalanced class data

As quality improvement efforts have been made in many applications such as healthcare and manufacturing systems, the number of adverse outcomes is gradually decreasing. The imbalanced class problem becomes very common where one class has significantly fewer samples than the other class. Most approaches have addressed the problem with focus on the improvement on the prediction of the minority class. In Chapter

IV, this research proposes two binary regression trees in the context of subgroup identification beyond prediction, Firth's logistic regression tree, and generalized extreme value regression tree. The potentials of the two proposed binary regression trees are investigated and compared with the logistic regression tree, which lays the foundation for developing effective subgroup identification models under the imbalanced class environment.

V.2 Future directions

This section describes three potential extensions of the dissertation work. The first direction is to improve the split variable selection step in the correlation tree in a statistically rigorous fashion. The second direction is to extend subgroup identification methods with diverse types of data and thus establish a subgroup surveillance scheme. The third one is to develop models that predict subgroups beyond identifying existing subgroups. Details of the future research are given as follows.

V.2.1 Correlation instability test for split variable selection

In the proposed correlation tree, the selected split variable is the one that has the largest p-value in the partial correlation test. However, the split variable selection test does not provide any information about how large the p-value is significantly meaningful. In other words, the test cannot rigorously evaluate how significantly the selected split variable explains the correlation of X_1 and X_2. In this case, the test is more likely to generate larger and complex tree since the test always splits unless the correlation of X_1 and X_2 is completely independent of the split variable.

This problem boils down to designing a rigorous hypothesis testing that investigates the association between correlation of X_1 and X_2 and the split variable Z_j. To handle this issue, we will propose another test for split variable selection, called correlation instability test. The hypothesis testing is formulated as follows.

$$\begin{aligned} H_0 &: Cor(\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}) \sim SN_2(\boldsymbol{0}, \rho) \quad for\ i = 1, \dots, n \\ H_1 &: Cor(\boldsymbol{X}_{1i}, \boldsymbol{X}_{2i}) \sim \begin{cases} SN_2(\boldsymbol{0}, \rho_1) & if\ \boldsymbol{X}_i \in R_1\,, \\ SN_2(\boldsymbol{0}, \rho_2) & if\ \boldsymbol{X}_i \in R_2 \end{cases} \end{aligned} \tag{V.1}$$

where SN_2 is the standard bivariate normal distribution. Essentially, the hypothesis testing is to conduct parameter instability test in the Chapter III over the standard bivariate normal density with parameter ρ. The test means that if the correlation of X_1 and X_2 is not statistically stable with respect to the covariate Z_j, the node *R* is split by the covariate Z_j. Given a significance level α, we can evaluate how significantly the covariate Z_j affects the correlation of X_1 and X_2, and thus decide whether the node *R* should be split or not.

V.2.2 Subgroup identification and surveillance for optimal personalized treatment

Through a subgroup identification model, my future research will lay the foundation for distinguishing groups of patients with different responses to treatments of interest. In particular, integrated with geospatial, text, image as well as clinical data, my future research aims to identify patients who are at a risk of diseases and need the more aggressive treatments from those who are less fitted to treatment because they already developed immunity or they will never progress. In a similar way, such personalized treatments can be extended to identify subpopulations of patients who can benefit more from a certain treatment over others. Furthermore, the current treatment plan at each subgroup will be monitored to check if it maintains the acceptable level of efficacy. In this way, we can improve patient care by administrating the right treatment for the right patients at the right time.

V.2.3 Real-time subgroup prediction model for effective control in healthcare

Beyond subgroup identification from historical observations, my future research aims to predict subgroups that have not spawned yet. Current subgroups detected from historical data can diverge into distinct subgroups or rather can converge into a single subgroup in the future. Such dynamics and uncertainty in subgrouping can be captured and modeled by virtualizing the current community and testing the influences of potential changes on the virtual community with continuously tracked health information and lifestyle parameters from the real world. The potential contribution is to find groups of patients and potential carriers who should be given the highest priority for preemptive intervention by observing the evolution of subgroups under infectious disease. Finally, such adaptive interventions will enable the health authority to make informed strategic decision and manage the spread of epidemic by identifying optimal resource allocations in advance.

REFERENCES

[1] Lipkovich, I., Dmitrienko, A., and B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, *36*(1), 136-196.

[2] Seibold, H., Zeileis, A., and Hothorn, T. (2016). Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics*, *12*(1), 45-63.

[3] Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, *30*(21), 2601-2621.

[4] Dusseldorp, E., Conversano, C., and Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: *STIMA. Journal of Computational and Graphical Statistics*, *19*(3), 514-530.

[5] Patel, S., Hee, S. W., Mistry, D., Jordan, J., Brown, S., Dritsaki, M., Ellard, D.R., Friede, T., Lamb, S.E., Lord, J., and Madan, J. (2016). Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials. *Programme Grants for Applied Research*, *4*(10).

[6] Chen, S., Tian, L., Cai, T., and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*(4), 1199-1209.

[7] Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867-2880.

[8] Xu, Y., Yu, M., Zhao, Y. Q., Li, Q., Wang, S., and Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics*, *71*(3), 645-653.

[9] Imai, K., and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, *7*(1), 443-470.

[10] Gu, X., Yin, G., and Lee, J. J. (2013). Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary Clinical Trials*, *36*(2), 642-650.

[11] Dusseldorp, E., and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine*, *33*(2), 219-237.

[12] Loh, W. Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, *34*(11), 1818-1833.

[13] Abend, G. (2017). What are neural correlates neural correlates of?. *BioSocieties*, *12*(3), 415-438.

[14] Dolcos, F., Iordan, A. D., and Dolcos, S. (2011). Neural correlates of emotion–cognition interactions: A review of evidence from brain imaging investigations. *Journal of Cognitive Psychology*, *23*(6), 669-694.

[15] Koch, C., Massimini, M., Boly, M., and Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307.

[16] Li, T., Luo, Q., and Gong, H. (2010). Gender-specific hemodynamics in prefrontal cortex during a verbal working memory task by near-infrared spectroscopy. *Behavioural Brain Research*, *209*(1), 148-153.

[17] Berchicci, M., Lucci, G., Perri, R. L., Spinelli, D., and Di Russo, F. (2014). Benefits of physical exercise on basic visuo-motor functions across age. *Frontiers in Aging Neuroscience*, *6*, 48.

[18] Davis, S. W., Dennis, N. A., Daselaar, S. M., Fleck, M. S., and Cabeza, R. (2007). Que PASA? The posterior–anterior shift in aging. *Cerebral Cortex*, *18*(5), 1201-1209.

[19] Abdullah, M. B. (1990). On a robust correlation coefficient. *The Statistician*, *39*(4), 455-460.

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

[21] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

[22] Chaudhuri, P., Huang, M. C., Loh, W. Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, *4*(1), 143-167.

[23] Loh, W. Y., and Shih, Y.S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*(4), 815–840.

[24] Shih, Y.S. (2004). A note on split selection bias in classification trees. *Computational Statistics and Data Analysis*, *45*(3), 457–466.

[25] Strobl, C., Boulesteix, A. L., and Augustin, T. (2007). Unbiased split selection for classification trees based on the Gini index. *Computational Statistics and Data Analysis*, *52*(1), 483–501.

[26] Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329-348.

[27] Levene, H. (1961). Robust tests for equality of variances. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, 279-292.

[28] Loh, W. Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica, 12*(2), 361-386.

[29] Wackerly, D., Mendenhall, W., and Scheaffer, R. L. (2014). *Mathematical statistics with applications.* Cengage Learning.

[30] Waliczek, T. M. (1996). *A primer on partial correlation coefficients.* Southwest Educational Research Association, New Orleans.

[31] Anderson, T. W. (2003). *An introduction to multivariate statistical analysis (3rd ed.).* Wiley, Hoboken, NJ.

[32] Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, *22*(6), 665.

[33] Whittaker, J. (2009). *Graphical models in applied multivariate statistics.* Wiley Publishing.

[34] Sun, S., Chen, J., Kind, P., Xu, L., Zhang, Y., and Burström, K. (2015). Experience-based VAS values for EQ-5D-3L health states in a national general population health survey in China, *Quality of Life Research. 24*(3), 693-703.

[35] Kirk, R. (2007) *Statistics: An introduction.* Nelson Education.

[36] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction (2nd ed.).* Springer.

[37] Xiao, W. (2017) An online algorithm for nonparametric correlations. *arXiv preprint arXiv:1712.01521*.

[38] Knight, W. R. (1966). A computer method for calculating kendall's tau with ungrouped data. *Journal of the American Statistical Association*, *61*(314), 436-439.

[39] Li, L., Cazzell, M., Zeng, L., and Liu, H. (2017). Are there gender differences in young vs. aging brains under risk decision-making? An optical brain imaging study, *Brain Imaging and Behaviour*. *11*(4), 1085-1098.

[40] Cazzell, M., Li, L., Lin, Z., Patel, S. J., and Liu, H. (2012). Comparison of neural correlates of risk decision making between genders: an exploratory fNIRS study of the balloon analogue risk task (BART). *Neuroimage*, *62*(3), 1896-1911.

[41] Therneau, T., Atkinson, B., and Ripley, B. (2015). rpart: Recursive Partitioning and Regression Trees. *R package version*, *4*, 1–9.

[42] Loh, W. Y. (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, *3*(4), 1710-1737.

[43] Rousselet, G. A., and Pernet, C. R. (2012). Improving standards in brain-behavior correlation analyses. *Frontiers in Human Neuroscience, 6*, 119.

[44] Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing (statistical modeling and decision science).* Academic Press.

[45] Hung, C., and Tsai, C. F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert Systems with Applications*, *34*(1), 780-787.

[46] Xu, L., Xu, Y., and Chow, T. W. (2010). PolSOM: A new method for multidimensional data visualization. *Pattern Recognition*, *43*(4), 1668-1675.

[47] Zhang, H., Wang, S., Xu, X., Chow, T. W., and Wu, Q. J. (2018). Tree2Vector: learning a vectorial representation for tree-structured data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(11), 5304-5318.

[48] Zhang, H., Wang, S., Zhao, M., Xu, X., and Ye, Y. (2018). Locality reconstruction models for book representation. *IEEE Transactions on Knowledge and Data Engineering*, *30*(10), 1873-1886.

[49] Kane, R. L., and Radosevich, D. M. (2010). *Conducting health outcomes research.* Jones & Bartlett Learning.

[50] Gerhardt, G., Yemane, A., Hickman, P., Oelschlaeger A., Rollins, E., and Brennan, N. (2013). Data shows reduction in medicare hospital readmission rates during 2012. *Medicare and Medicaid Research Review*, *3*(2), E1-E11.

[51] Dawson J., Doll, H., Fitzpatrick, R., Jenkinson, C., and Carr, A. J. (2010). Routine use of patient reported outcome measures in healthcare settings. *British Medical Journal*, *340*, 464-467.

[52] Knaup, C., Koesters, M., Schorfer, D., Becker, T., and Puschner, B. (2009). Effect of feedback of treatment outcome in specialist mental healthcare: meta-analysis. *The British Journal of Psychiatry*, *195*(1), 15-22.

[53] Complex Systems Modeling Group. (2010). *Modeling in healthcare.* American Mathematical Society.

[54] Yamashita, T., Bailer, A. J., and Noe, D. A. (2013). Identifying at-risk subpopulations of Canadians with limited health literacy. *Epidemiology Research International*, Article ID 130263.

[55] Zeng, L., Neogi, S., Rogers, J., Seidensticker, S., Clark, C., Sonstein, L., Trevino, R., and Sharma, G. (2014). Statistical Models for Hospital Readmission Prediction with Application to Chronic Obstructive Pulmonary Disease (COPD) Patients. *Proceedings of the 4th International Conference on Industrial Engineering and Operations Management (IEOM)*, Bali, Indonesia.

[56] Yamashita, T., Kart C. S., and Noe, D. A. (2012). Predictors of adherence with self-care guidelines among persons with type 2 diabetes: results from a logistic regression tree analysis. *Journal of Behavioral Medicine*, *35*(6), 603-615.

[57] Yamashita, T., Noe, D. A., and Bailer, A. J. (2012). Risk factors of falls in community-dwelling older adults: logistic regression tree analysis. *The Gerontologist*, *52*(6), 822-832.

[58] Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

[59] Mohammed, M. A., and Laney, D. (2006). Overdispersion in health care performance data: Laney’s approach. *Quality and Safety in Healthcare*, *15*(5), 383-384.

[60] Yu, C., and Yao, W. (2017). Robust linear regression: a review and comparison. *Communications in Statistics-Simulation and Computation*. *46*(8), 6261-6282.

[61] Nurunnabi, A., and West, G. (2012). Outlier detection in logistic regression: a quest for reliable knowledge from predictive modeling and classification. *In 2012 IEEE 12th International Conference on Data Mining Workshops*, 643-652.

[62] Cordeiro, G. M. (2004). On Pearson's residuals in generalized linear models. *Statistics & probability letters*, *66*(3), 213-219.

[63] Gelman, A., and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

[64] Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492-514.

[65] Chaudhuri, P., Lo, W. D., Loh, W. Y., and Yang, C. C. (1995). Generalized regression trees. *Statistica Sinica*, *5*(2), 641-666.

[66] Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika*, *74*(3), 503-515.

[67] Chan, K. Y., and Loh, W. Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, *13*(4), 826-852.

[68] Doyle, P. (1973). The use of automatic interaction detector and similar search procedures. *Journal of the Operational Research Society*, *24*(3), 465-467.

[69] Landwehr, N., Hall, M., and Frank, E. (2005). Logistic model trees. *Machine learning*, *59*(1-2), 161-205.

[70] Quinlan, J. R. (1993). C4.5: Programming for machine learning. *Morgan Kauffmann*, *38*, 48.

[71] Lee, S., and Jun, C. H. (2018). Fast incremental learning of logistic model tree using least angle regression. *Expert Systems with Applications*, *97*, 137-145.

[72] Zeileis, A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, *24*(4), 445-466.

[73] Zeileis, A., and Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488-508.

[74] Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, *61*(4), 821-856.

[75] Hansen, B. E. (1997). Approximate asymptotic p values for structural-change tests. *Journal of Business & Economic Statistics*, *15*(1), 60-67.

[76] Hjort, N. L., and Koning, A. (2002). Tests for constancy of model parameters over time. *Journal of Nonparametric Statistics*, *14*(1-2), 113-132.

[77] Zhang, Z. (2016). Residuals and regression diagnostics: focusing on logistic regression. *Annals of Translational Medicine*, *4*(10), 195.

[78] Imon, A. R., and Hadi, A. S. (2008). Identification of multiple outliers in logistic regression. *Communications in Statistics-Theory and Methods*, *37*(11), 1697-1709.

[79] Jennings, D. E. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association*, *81*(396), 987-990.

[80] Cordeiro, G. M., and Simas, A. B. (2009). The distribution of Pearson residuals in generalized linear models. *Computational Statistics & Data Analysis, 53*(9), 3397-3411.

[81] Duffy, D. E. (1990). On continuity-corrected residuals in logistic regression. *Biometrika*, *77*(2), 287-293.

[82] Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, *38*(2), 485-498.

[83] Stefanski, L. A., Carroll, R. J., and Ruppert, D. (1986). Optimally hounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, *73*(2), 413-424.

[84] Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989). Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, *84*(406), 460-466.

[85] Morgenthaler, S. (1992). Least-absolute-deviations fits for generalized linear models. *Biometrika*, *79*(4), 747-754.

[86] Croux, C., and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis*, *44*(1-2), 273-295.

[87] Hobza, T., Pardo, L., and Vajda, I. (2008). Robust median estimator in logistic regression. *Journal of Statistical Planning and Inference*, *138*(12), 3822-3840.

[88] Park, S. Y., and Liu, Y. (2011). Robust penalized logistic regression with truncated loss functions. *Canadian Journal of Statistics*, *39*(2), 300-323.

[89] Park, H., and Konishi, S. (2016). Robust logistic regression modelling via the elastic net-type regularization and tuning parameter selection. *Journal of Statistical Computation and Simulation*, *86*(7), 1450-1461.

[90] Feng, J., Xu, H., Mannor, S., and Yan, S. (2014). Robust logistic regression and classification. *In Advances in Neural Information Processing Systems*, 253-261.

[91] Barnett, V. and Lewis, T. (1994). *Outliers in statistical data (3rd ed.)*. John Wiley & Sons.

[92] Rousseeuw, P. J., and Leroy, A. M. (2005). *Robust regression and outlier detection*. John Wiley & Sons.

[93] Heritier, S., Cantoni, E., Copt, S., and Victoria-Feser, M. P. (2009). *Robust methods in biostatistics*. John Wiley & Sons.

[94] Albert, J., and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, *82*(4), 747-769.

[95] Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, *88*(422), 669-679.

[96] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

[97] Global Initiative for Chronic Obstructive Lung Disease (2019). *Global strategy for the diagnosis, management, and prevention of COPD, 2020 Report*. http://goldcopd.org/gold-reports.

[98] Sonstein, L., Clark, C., Seidensticker, S., Zeng, L., and Sharma, G. (2014). Improving adherence for management of acute exacerbation of chronic obstructive pulmonary disease. *The American Journal of Medicine*, *127*(11), 1097-1104.

[99] Jones, K. W., Jackson, M., Grotte, G., and Bridgewater, B. (2000). Limitations of the parsonnet score for measuring risk stratified mortality in the north west of England. *Heart*, *84*(1), 71-78.

[100] Steiner, S. H., Cook, R. J., Farewell, V. T., and Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, *1*(4), 441-452.

[101] Sego, L. H., Reynolds Jr, M. R., and Woodall, W. H. (2009). Risk-adjusted monitoring of survival times. *Statistics in Medicine*, *28*(9), 1386-1401.

[102] Steiner, S. H., and Jones, M. (2010). Risk-adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. *Statistics in Medicine, 29*(4), 444-454.

[103] Zeng, L., and Zhou, S. (2011). A Bayesian approach to risk-adjusted outcome monitoring in healthcare. *Statistics in Medicine, 30*(29), 3431-3446.

[104] Paynabar, K., Jin, J., and Yeh, A. B. (2012). Phase I risk-adjusted control charts for monitoring surgical performance by considering categorical covariates. *Journal of Quality Technology*, *44*(1), 39-53.

[105] Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339-1349.

[106] D Sun, Y., Kamel, M. S., Wong, A. K., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358-3378.

[107] Elkan, C. (2001). The foundations of cost-sensitive learning. *In International Joint Conference on Artificial Intelligence*, *17*(1), 973-978.

[108] Japkowicz, N., and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429-449.

[109] Byon, E., Shrivastava, A. K., and Ding, Y. (2010). A classification procedure for highly imbalanced class sizes. *IIE Transactions*, *42*(4), 288-303.

[110] Pourhabib, A., Mallick, B. K., and Ding, Y. (2015). Absent data generating classifier for imbalanced class sizes. *Journal of Machine Learning Research, 16*(1), 2695-2724.

[111] Park, C., Huang, J. Z., and Ding, Y. (2010). A computable plug-in estimator of minimum volume sets for novelty detection. *Operations Research*, *58*(5), 1469-1480.

[112] López, V., Fernández, A., Moreno-Torres, J. G., and Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications, 39*(7), 6585-6608.

[113] Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. *In Inter-2003 Workshop on Learning from Imbalanced Data sets II, 2*.

[114] Drummond, C., and Holte, R. C. (2000). Exploiting the cost (in) sensitivity of decision tree splitting criteria. *In Proceedings of the Seventeenth International Conference on Machine Learning*, *1*(1), 239-246.

[115] Dietterich, T., Kearns, M., and Mansour, Y. (1996). Applying the weak learning framework to understand and improve C4.5. *In Proceedings of the Thirteenth International Conference on Machine Learning,* 96-104.

[116] Liu, W., Chawla, S., Cieslak, D. A., and Chawla, N. V. (2010). A robust decision tree algorithm for imbalanced data sets. *In Proceedings of the 2010 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics*, 766-777.

[117] Cieslak, D. A., Hoens, T. R., Chawla, N. V., and Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery, 24*(1), 136-158.

[118] Maszczyk, T., and Duch, W. (2008). Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. *In International Conference on Artificial Intelligence and Soft Computing*, 643-651.

[119] Park, Y., and Ghosh, J. (2012). Ensembles of (α)-Trees for Imbalanced Classification Problems. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 131-143.

[120] Hong, C., Ghosh, R., and Srinivasan, S. (2016). Dealing with class imbalance using thresholding. *arXiv preprint arXiv:1607.02705*.

[121] Casella, G., and Berger, R. L. (2002). *Statistical inference*, Pacific Grove, CA: Duxbury, 337-472.

[122] King, G., and Zeng, L. (2001). Logistic regression in rare events *data*. *Political Analysis*, *9*(2), 137-163.

[123] Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 27-38.

[124] Heinze, G., and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, *21*(16), 2409-2419.

[125] Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T. G., Altamirano, A., and Yaitul, V. (2018). A study on the effects of unbalanced data when fitting logistic regression models in ecology. *Ecological Indicators*, *85*, 502-508.

[126] Maalouf, M., Homouz, D., and Trafalis, T. B. (2018). Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. *Computational Intelligence, 34*(1), 161-174.

[127] Oommen, T., Baise, L. G., and Vogel, R. M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Mathematical Geosciences*, *43*(1), 99-120.

[128] Owen, A. B. (2007). Infinitely imbalanced logistic regression. *Journal of Machine Learning Research*, *8*, 761-773.

[129] Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society*. *Series B (Methodological)*, 310-313.

[130] Rahayu, S. P. (2012). *Logistic Regression Methods for Classification of Imbalanced Data Sets* (Doctoral dissertation, UMP).

[131] Maalouf, M., and Siddiqi, M. (2014). Weighted logistic regression for large-scale imbalanced and rare events data. *Knowledge-Based Systems*, *59*, 142-148.

[132] Komarek, P., and Moore, A. W. (2005). Making logistic regression a core data mining tool with tr-irls. *In Fifth IEEE International Conference on Data Mining*.

[133] Maalouf, M., and Trafalis, T. B. (2011). Robust weighted kernel logistic regression in imbalanced and rare events data. *Computational Statistics & Data Analysis*, *55*(1), 168-183.

[134] Wang, H., Xu, Q., and Zhou, L. (2015). Large unbalanced credit scoring using lasso-logistic regression ensemble. *PloS one*, *10*(2).

[135] McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.).* Edition Chapman and Hall. London, UK.

[136] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*(1007), 453-461.

[137] Cole, S. R., Chu, H., and Greenland, S. (2014). Maximum likelihood, profile likelihood, and penalized likelihood: a primer. *American Journal of Epidemiology*, *179*(2), 252-260.

[138] King, E. N., and Ryan, T. P. (2002). A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician*, *56*(3), 163-170.

[139] Calabrese, R., and Osmetti, S. A. (2013). Modelling small and medium enterprise loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, *40*(6), 1172-1188.

[140] Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, *72*(1), 67-90.

[141] El Adlouni, S., Ouarda, T. B., Zhang, X., Roy, R., & Bobée, B. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, *43*(3).

[142] Cheng, R. C. H., and Iles, T. C. (1987). Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *49*(1), 95-101.

[143] Davison, A. C., and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(3), 393-425.

[144] Huard, D., Mailhot, A., and Duchesne, S. (2010). Bayesian estimation of intensity–duration–frequency curves and of the return period associated to a given rainfall event. *Stochastic Environmental Research and Risk Assessment*, *24*(3), 337-347.

[145] Lee, Y., Shin, Y., and Park, J. S. (2017). A data-adaptive maximum penalized likelihood estimation for the generalized extreme value distribution. *Communications for Statistical Applications and Methods*, *24*(5), 493-505.

[146] Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, *27*(3), 251-261.

[147] Wang, X., and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*, *4*(4), 2000-2023.

[148] Agarwal, A., Narasimhan, H., Kalyanakrishnan, S., and Agarwal, S. (2014). Gev-canonical regression for accurate binary class probability estimation when one class is rare. *In International Conference on Machine Learning*, 1989-1997.

[149] Zhang, H., Liu, G., Pan, L., Meng, K., and Li, J. (2016). GEV regression with convex loss applied to imbalanced binary classification. *In 2016 IEEE First International Conference on Data Science in Cyberspace,* 532-537.

APPENDIX A

SUPPLEMENTAL MATERIALS FOR CHAPTER III

A.1 Derivation of Equations (III.11) and (III.12)

The following is a well-known result on generalized linear models based on exponential family density:

$$f(Y_i|\theta_i,\phi) = exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \varphi(Y_i,\phi)\right\},$$

where θ_i is the canonical parameter, the functions b and φ are specific to individual distribution, and ϕ is the dispersion parameter common to all Y_i. Given the log-likelihood function, $l(Y_i|\theta_i,\phi) = \log f(Y_i|\theta_i,\phi)$, and the facts $\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0$ and $\mathrm{Var}\left(\frac{\partial l}{\partial \theta}\right) = -\mathrm{E}\left(\frac{\partial^2 l}{\partial \theta^2}\right)$, we have

$$\mathrm{E}(Y_i) = \mu_i = b'(\theta_i),$$

$$\mathrm{Var}(Y_i) = a_i(\phi)b''(\theta_i) = a_i(\phi)\mathrm{Var}(\mu_i) = a_i(\phi)V(\mu_i).$$

In addition, $\eta_i = g(\mu_i) = \boldsymbol{X}_i^T\boldsymbol{\beta},$ where g is the known link function i.e., $\mu_i = g^{-1}(\boldsymbol{X}_i^T\boldsymbol{\beta})$. Then, the k^{th} element of the score function in the generalized linear model is

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n}\frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_k} = \sum_{i=1}^{n}\frac{1}{a_i(\phi)V(\mu_i)g'(\mu_i)}(Y_i - \mu_i)X_{ik}.$$

Logistic regression is a special case of the exponential family under the following setting

$$\theta_i = \log\frac{\mu_i}{1-\mu_i}, \qquad b(\theta_i) = log(1-\mu_i),$$

$$\varphi(Y_i,\phi) = log\binom{1}{Y_i},\ g(\mu_i) = log\frac{\mu_i}{1-\mu_i}, a_i(\phi) = 1.$$

Based on the above functions and weights w_i in Eq. (10), the k^{th} element of the robust score function $\boldsymbol{s}_{robust}\left(Y_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}_{robust}\right)$ is given by

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} \frac{w_i}{V(\mu_i)g'(\mu_i)}(Y_i - \mu_i)X_{ik} = \sum_{i=1}^{n} w_i\left(Y_i - g^{-1}\left(\boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{robust}\right)\right)X_{ik}$$

which yields Equation (III. 11). For the robust version of covariance matrix in Equation (III.12), it is obtained by plugging the robust score function $\boldsymbol{s}_{robust}$ into the OPG estimator $\widehat{\boldsymbol{J}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}\boldsymbol{s}^{\boldsymbol{T}}$.

A.2 Estimation of posterior distribution of latent residual

The posterior distribution of the latent residual $P(\varepsilon_i|\mathcal{D})$ is obtained by calculating $\varepsilon_i = \xi_i - \boldsymbol{X}_i^{\boldsymbol{T}}\boldsymbol{\beta}$ for a number of samples of ξ_i and $\boldsymbol{\beta}$ from their joint posterior distribution [94]. The joint posterior based on the *t*-link function is written as [95]

$$p\left(\xi, \boldsymbol{\beta}, \lambda, v\middle|\mathcal{D}\right) \propto p(v)\prod_{i=1}^{n}\{1(\xi_i > 0)1(Y_i = 1) + 1(\xi_i \leq 0)1(Y_i = 0)\}\sqrt{\lambda_i/2\pi}$$
$$\times e^{\left(\frac{-\lambda_i}{2\left(\xi_i - x_i^T\boldsymbol{\beta}\right)^2}\right)} c(v)\lambda_i^{\frac{v}{2}-1} e^{-v\lambda_i/2},$$

where v is the degrees of freedom of the *t*-link function, $p(v)$ is the prior of v, λ_i is an additional parameter to represent the *t* distribution as a scale mixture of normal distributions, and $c(v)$ is$\left[\Gamma(v/2)\left((v/2)\right)^{(v/2)}\right]^{-1}$. Since v is set to 8, we only have three unknown parameters, ξ, $\boldsymbol{\beta}$, and λ . Given a uniform prior for $\boldsymbol{\beta}$, the joint posterior of the three parameters can be obtained by Gibbs sampling that iteratively draws samples from the conditional posteriors as follows:

$$Step\ 1.\quad p\left(\xi_i \middle| \mathcal{D}, \boldsymbol{\beta}\ , \lambda\right) \sim N\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\ , \lambda_i^{-1}\right)$$

$$\begin{matrix} truncated\ at\ the\ left\ by\ 0 & if\ Y_i = 1 \\ truncated\ at\ the\ rigft\ by\ 0 & if\ Y_i = 0 \end{matrix}\quad i = 1, \ldots, n$$

$$Step\ 2.\quad \begin{matrix} p(\boldsymbol{\beta}|\mathcal{D}, \xi, \lambda) \sim N\left(\widehat{\boldsymbol{\beta}}_{\xi,\lambda},\ (\boldsymbol{X^T W X})^{-1}\right) \\ \text{where } \widehat{\boldsymbol{\beta}}_{\xi,\lambda} = (\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W}\xi \text{ and } \boldsymbol{W} = diag(\lambda_i) \end{matrix}$$

$$Step\ 3.\quad p(\lambda_i|\mathcal{D}, \xi_i, \boldsymbol{\beta}) \sim Gamma\left(\frac{9}{2},\ \frac{2}{8+\left(\xi_i - \boldsymbol{X}_i^T\boldsymbol{\beta}\right)^2}\right)$$

The maximum likelihood estimates of $\boldsymbol{\beta}$ from a standard logistic regression and $\lambda_i = 1$, $i = 1, \ldots, n$, can be used as initial values in Step 1.

A.3 Algorithm of the proposed method

Algorithm 1.

```
1.  Initialize $K$ = 2.306
2.  Identify outliers in each class of $Y$
        (a) Compute $M_i$, $i = 1, \dots, n$
        (b) Sort $M_i$s in descending order to obtain $M_{(i)}$, $i = 1, \dots, n$
        (c) Compute adjacent differences $Chasm_i$ of $M_{(i)}$, $i = 1, \dots, n$
        (d) Find the location $i^* = \max_{1 \le i \le n-1} Chasm_i$
        (e) Take original observations corresponding to $\{M_{(1)}, \dots, M_{(i^*)}\}$ as outliers
3.  Conduct parameter instability test
        (a) For $r = L, L + 1, \dots, U$:
            (a1) Draw $r$ samples randomly from the outlier set and return them to the
                 normal set
            (a2) Calculate the original test statistic $\mathcal{T}$ using the updated normal samples
            (a3) Repeat Steps (a1) and (a2) $m$ times
            (a4) Calculate average of the test statistics $\mathcal{T}_r^{ave}$
        (b) Find the maximum among $\{\mathcal{T}_L^{ave}, \mathcal{T}_{L+1}^{ave}, \dots, \mathcal{T}_U^{ave}\}$ as the overall test statistic
        (c) Calculate the p-value of the test statistic
        (d) Split the node if p-value < 0.05 and not split otherwise
```

APPENDIX B

SUPPLEMENTAL MATERIALS FOR CHAPTER IV

B.1 Fisher information matrix for parameters of GEV regression

The Fisher information is the negative of the expectation of the second derivatives of the log-likelihood with respect to parameters. Using the chain rule of higher order partial derivatives, the second order partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$ and τ are defined as below

$$\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial^2\beta_j} = \sum_{i=1}^{n} \frac{\partial^2 l_i(\pi(\boldsymbol{X}_i)\,)}{\partial^2\pi(\boldsymbol{X}_i)} \left[\frac{\partial\pi(\boldsymbol{X}_i)}{\partial\beta_j}\right]^2 + \frac{\partial l_i(\pi(\boldsymbol{X}_i)\,)}{\partial\pi(\boldsymbol{X}_i)} \frac{\partial^2\pi(\boldsymbol{X}_i)}{\partial^2\beta_j},$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial^2\tau} = \sum_{i=1}^{n} \frac{\partial^2 l_i(\pi(\boldsymbol{X}_i)\,)}{\partial^2\pi(\boldsymbol{X}_i)} \left[\frac{\partial\pi(\boldsymbol{X}_i)}{\partial\tau}\right]^2 + \frac{\partial l_i(\pi(\boldsymbol{X}_i)\,)}{\partial\pi(\boldsymbol{X}_i)} \frac{\partial^2\pi(\boldsymbol{X}_i)}{\partial^2\tau},$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial^2\beta_j\beta_k} = \sum_{i=1}^{n} \frac{\partial^2 l_i(\pi(\boldsymbol{X}_i)\,)}{\partial^2\pi(\boldsymbol{X}_i)} \frac{\partial\pi(\boldsymbol{X}_i)}{\partial\beta_j} \frac{\partial\pi(\boldsymbol{X}_i)}{\partial\beta_k} + \frac{\partial l_i(\pi(\boldsymbol{X}_i)\,)}{\partial\pi(\boldsymbol{X}_i)} \frac{\partial^2\pi(\boldsymbol{X}_i)}{\partial^2\beta_j\beta_k},$$

$$\frac{\partial^2 l(\boldsymbol{\beta},\tau,Y_i\,)}{\partial\beta_j\partial\tau} = \sum_{i=1}^{n} \frac{\partial}{\partial\beta_j}\left[\frac{\partial l_i(\boldsymbol{\beta},\tau)}{\partial\tau}\right],$$

where

$$\frac{\partial^2 l_i(\pi(\boldsymbol{X}_i)\,)}{\partial^2\pi(\boldsymbol{X}_i)} = -\frac{Y_i}{[\pi(\boldsymbol{X}_i)]^2} - \frac{1-Y_i}{[1-\pi(\boldsymbol{X}_i)]^2},$$

$$\frac{\partial^2(\pi(\boldsymbol{X}_i)\,)}{\partial^2\beta_j\beta_k} = X_{ij}X_{ik}\pi(\boldsymbol{X}_i)(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta})^{-\frac{1}{\tau}-2}\{1+\tau+\ln[\pi(\boldsymbol{X}_i)]\}$$

$$\frac{\partial^2(\pi(\boldsymbol{X}_i)\,)}{\partial^2\tau} = \pi(\boldsymbol{X}_i)\ln[\pi(\boldsymbol{X}_i)]\{B_1+B_2\}$$

where $B_1 = \left[\frac{1}{\tau^2}\ln(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta}) - \frac{\boldsymbol{X}_i^T\boldsymbol{\beta}}{\tau\left(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta}\right)}\right]^2 [\ln[\pi(\boldsymbol{X}_i)]+1]$ and

$$B_2 = \left[-\frac{2}{\tau^3}\ln(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta}) + \frac{\boldsymbol{X}_i^T\boldsymbol{\beta}+2\tau\left(\boldsymbol{X}_i^T\boldsymbol{\beta}\right)^2+\boldsymbol{X}_i^T\boldsymbol{\beta}\left(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta}\right)}{\tau^2\left(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta}\right)^2}\right].$$

Based on the second order partial derivatives and $E\left(\frac{\partial^2 l_i(\pi(\boldsymbol{X}_i)\,)}{\partial\pi(\boldsymbol{X}_i)}\right) = 0$, Fisher information is given as

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial^2\beta_j}\right) = -\sum_{i=1}^{n}\frac{1}{\pi(\boldsymbol{X}_i)[1-\pi(\boldsymbol{X}_i)]}\left[\frac{\partial l(\boldsymbol{\beta},\tau)}{\partial\beta_j}\right]^2$$

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial^2\tau}\right) = -\sum_{i=1}^{n}\frac{\partial^2(\pi(\boldsymbol{X}_i)\,)}{\partial^2\tau}\frac{1}{\pi(\boldsymbol{X}_i)[1-\pi(\boldsymbol{X}_i)]}$$

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial\beta_j\partial\beta_k}\right) = -\sum_{i=1}^{n}\frac{1}{\pi(\boldsymbol{X}_i)[1-\pi(\boldsymbol{X}_i)]}\frac{\partial l(\boldsymbol{\beta},\tau)}{\partial\beta_j}\frac{\partial l(\boldsymbol{\beta},\tau)}{\partial\beta_k}$$

$$-E\left(\frac{\partial^2 l(\boldsymbol{\beta},\tau)}{\partial\beta_j\partial\tau}\right) = -\sum_{i=1}^{n}X_{ij}\frac{\ln^2[\pi(\boldsymbol{X}_i)]\pi(\boldsymbol{X}_i)}{(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta})[1-\pi(\boldsymbol{X}_i)]}\left[\frac{1}{\tau^2}\ln(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta}) - \frac{\boldsymbol{X}_i^T\boldsymbol{\beta}}{\tau(1+\tau\boldsymbol{X}_i^T\boldsymbol{\beta})}\right]$$

B.2 The values of intercept used in the change-case in the simulation

			Degree of change					
			5 %	10 %	20 %	40 %	60 %	80 %
Degree of balance	1 %	$\boldsymbol{\theta}_1$	−5.04	−5.00	−5.03	−5.06	−5.32	−5.58
		$\boldsymbol{\theta}_2$	−5.02	−4.89	−4.81	−4.48	−4.45	−4.38
	5 %	$\boldsymbol{\theta}_1$	−3.51	−3.58	−3.78	−3.87	−4.18	−4.92
		$\boldsymbol{\theta}_2$	−3.35	−3.31	−3.22	−3.16	−3.10	−3.01
	10 %	$\boldsymbol{\theta}_1$	−2.79	−3.04	−3.10	−3.22	−3.78	−4.18
		$\boldsymbol{\theta}_2$	−2.83	−2.80	−2.71	−2.58	−2.47	−2.37
	15%	$\boldsymbol{\theta}_1$	−2.44	−2.55	−2.71	−3.01	−3.22	−3.87
		$\boldsymbol{\theta}_2$	−2.41	−2.40	−2.37	−2.09	−1.95	−1.83
	20%	$\boldsymbol{\theta}_1$	−2.26	−2.30	−2.39	−2.60	−3.04	−3.72
		$\boldsymbol{\theta}_2$	−2.17	−2.03	−1.93	−1.85	−1.69	−1.52
	25%	$\boldsymbol{\theta}_1$	−1.95	−1.96	−2.23	−2.50	−2.87	−3.55
		$\boldsymbol{\theta}_2$	−1.94	−1.88	−1.73	−1.57	−1.32	−1.25
	30 %	$\boldsymbol{\theta}_1$	−1.83	−1.84	−1.93	−2.30	−2.60	−3.35
		$\boldsymbol{\theta}_2$	−1.65	−1.63	−1.52	−1.34	−1.18	−0.95
	35 %	$\boldsymbol{\theta}_1$	−1.56	−1.65	−1.85	−2.15	−2.62	−3.07
		$\boldsymbol{\theta}_2$	−1.54	−1.48	−1.34	−1.16	−1.01	−0.86
	40 %	$\boldsymbol{\theta}_1$	−1.45	−1.56	−1.69	−1.93	−2.39	−3.04
		$\boldsymbol{\theta}_2$	−1.36	−1.26	−1.18	−1.01	−0.85	−0.65
	45 %	$\boldsymbol{\theta}_1$	−1.35	−1.43	−1.52	−1.88	−2.30	−2.96
		$\boldsymbol{\theta}_2$	−1.13	−1.09	−0.95	−0.86	−0.65	−0.43