Machine Learning Models for High-Dimensional Matched Data

by

Nooshin Shomal Zadeh

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2021 by the
Graduate Supervisory Committee:

George C. Runger, Chair
Douglas Montgomery
Adolfo R. Escobedo
Shilpa Shinde

ARIZONA STATE UNIVERSITY

August 2021

ABSTRACT

Matching or stratification is commonly used in observational studies to remove bias due to confounding variables. Analyzing matched data sets requires specific methods which handle dependency among observations within a stratum. Also, modern studies often include hundreds or thousands of variables. Traditional methods for matched data sets are challenged in high-dimensional settings, mixed type variables (numerical and categorical), nonlinear and interaction effects. Furthermore, machine learning research for such structured data is quite limited. This dissertation addresses this important gap and proposes machine learning models for identifying informative variables from high-dimensional matched data sets.

The first part of this dissertation proposes a machine learning model to identify informative variables from high-dimensional matched case-control data sets. The outcome of interest in this study design is binary (case or control), and each stratum is assumed to have one unit from each outcome level. The proposed method which is referred to as Matched Forest (MF) is effective for large number of variables and identifying interaction effects.

The second part of this dissertation proposes three enhancements of MF algorithm. First, a regularization framework is proposed to improve variable selection performance in excessively high-dimensional settings. Second, a classification method is proposed to classify unlabeled pairs of data. Third, two metrics are proposed to estimate the effects of important variables identified by MF.

The third part proposes a machine learning model based on Neural Networks to identify important variables from a more generalized matched case-control data set where each stratum has one unit from case outcome level and more than one unit from control outcome level. This method which is referred to as Matched Neural

Network (MNN) performs better than current algorithms to identify variables with interaction effects.

Lastly, a generalized machine learning model is proposed to identify informative variables from high-dimensional matched data sets where the outcome has more than two levels. This method outperforms existing algorithms in the literature in identifying variables with complex nonlinear and interaction effects.

DEDICATION

I dedicate my work to my dearest parents and my beloved husband for their constant support and encouragement over this journey.

ACKNOWLEDGMENTS

Throughout my Ph.D. study, I have received a great deal of support and assistance. I am happy to take this opportunity to thank all the people who have supported me in this journey.

I would like to express my sincere gratitude to my advisor Dr. George Runger for his invaluable advice, continuous support, and patience during my Ph.D. study. His insightful feedback on my research taught me how to look at real world problems from a scientific point of view. This accomplishment would have not been possible without his guidance.

I would like to extend my sincere thanks to my dissertation committee members Dr. Douglas Montgomery, Dr. Adolfo Escobedo, and Dr. Shilpa Shinde. Without their support and invaluable feedback on my research, this dissertation would have never been accomplished.

Finally, I would like to thank my dearest friends Ghazal Shams, Fatemeh Karamyar, Sangdi Lin, Maziar kasaei and Yonghui Fan and many others who have supported me throughout this journey in many aspects, including but not limited to research, life and job. I am sincerely grateful for all their advice, encouragement, and the friendship that we have built.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1 INTRODUCTION 1

2 BACKGROUND 7

2.1 Matched Case-Control Study Design 7

2.2 Variable Selection Methods For Matched Case-Control Data Sets ... 8

2.2.1 Conditional Logistic Regression and Variants for High-dimensional Setting 8

2.2.2 Generalized Linear Mixed Models 12

2.2.3 Boosting Methods 12

2.2.4 Conditional Classification Algorithms 13

2.3 Variable Importance Measure Via Random Forest 14

2.4 Variable Importance Measure Via Artificial Neural Networks 16

3 MATCHED FOREST: SUPERVISED LEARNING FOR HIGH-DIMENSIONAL MATCHED CASE-CONTROL STUDIES 21

3.1 Introduction 21

3.2 Background 23

3.2.1 Matched Case-Control Analysis 23

3.2.2 Existing Methods For High-Dimensional Matched Case-Control Data Sets 25

3.3 Matched Forest 28

3.3.1 Transform to Supervised Learning 28

3.3.2 MF Variable Importance 32

CHAPTER Page

3.4 Experiments 33
3.4.1 Simulation Studies 34
3.4.2 Biomedical Examples 36
3.5 Conclusions 43
3.6 Supporting Information: Simulation Studies 44
3.6.1 Null Scenario 44
3.6.2 Linear Exposure Effect 45
3.6.3 Non-Linear Exposure Effect 46
3.6.4 Matching-Exposure Interaction 47
3.6.5 Exposure-Exposure Interaction 48
3.6.6 Exposure-Exposure-Exposure Interaction 49
3.6.7 Summary of simulation results 52
4 ENHANCEMENTS OF MATCHED FOREST 60
4.1 Introduction 60
4.2 Background 62
4.2.1 Classification of Matched Case-Control Data Sets 62
4.2.2 Matched Forest (MF) 63
4.3 Weighted Matched Forest, Classification And Effect Estimation 65
4.3.1 Weighted Matched Forest 65
4.3.2 Matched Forest for Classification of Matched Pairs 67
4.3.3 Effect Estimation 69
4.4 Experiments 71
4.4.1 Simulations 71
4.4.2 Case Studies 87

CHAPTER Page

4.5 Conclusion 92
5 MATCHED CASE-CONTROL ANALYSIS USING NEURAL NETWORKS 93
5.1 Background 95
5.1.1 Variable Selection Methods For High-dimensional Matched Case-control Datasets 95
5.1.2 DeepSHAP 96
5.2 Method: Matched Neural Network 99
5.2.1 Matched $1-1$ Study Design 100
5.2.2 Matched $1-L$ Study Design 102
5.3 Experiments 104
5.3.1 Simulation Study 104
5.4 Case Study: Childhood Acute Lymphoblastic Leukemia Study 111
5.5 Conclusion 113
6 MULTINOMIAL MATCHED LEARNER: SUPERVISED MACHINE LEARNING FOR HIGH-DIMENSIONAL MATCHED STUDIES WITH MULTIPLE LEVELS OF THE OUTCOME 118
6.1 Introduction 118
6.2 Background 121
6.2.1 Related Work 121
6.2.2 Matched Forest 122
6.3 Multinomial Matched Learner 125
6.3.1 Data Transformation To Supervised Setting 126
6.3.2 Variable Importance 131

CHAPTER Page

6.4 Experiments 133
6.4.1 Simulation Study 133
6.4.2 Clinical Application 137
6.5 Conclusion 140
7 CONCLUSION 142
REFERENCES 144

LIST OF TABLES

Table Page

3.1 Design Table for Exposure-exposure-exposure Interaction Simulation. . . 49

4.1 Prediction of a New Pair Based on the Predicted Labels for Its Observed and Counterfactual Pairs 68

LIST OF FIGURES

Figure Page

3.1 Scatter Plot of Control (x^0) Versus Case (x^1) for Exposure Variable x with (a) No Effect and (b) Effect 24

3.2 Scatter Plot of Control Versus Case for An Exposure Variable x with Matching-exposure Interaction Effect. 25

3.3 Scatter Plot of New Control (x^{*0}) Versus New Case (x^{*1}) Variables Associated with Exposure Variable x with (a) No Effect and (B) An Effect. 31

3.4 Scatter Plot of Control x^{*0} Versus Case x^{*1} (a) Without and (b) with Matching Variable v Illustrate An Interaction Effect Between x and v. . 32

3.5 Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy for Data Sets Simulated With Matching-Exposure Interaction 37

3.6 Indian Liver Patient Data Set: (a) MFI Scores for Exposure Variables and (b) P-values from MF and CLR. 39

3.7 Pima Indians Diabetes Data Set: (a) MFI Scores for Exposure Variables and (b) P-values from MF and CLR. 40

3.8 Statlog Heart Disease Data Set: (a) MFI Scores for Exposure Variables and (b) P-values from MF and CLR. 41

3.9 (a) Average MFI Scores for the 100 Highest MFI Importance Variables. Genes Selected by WL_2Boost Are Shown in Dark Shade. (b) Scatter Plot of Control Versus Case for Gene 213166_x_at Indicates An Effect. . 41

3.10 Null Scenario With Independent Variables: (a) Comparison Between the Performance of MF, RF, and CLR in Terms of FPR, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated with $N = 600$ and $R = 100$. MF and CLR Have Small and Similar FPR and If $\alpha < 0.3$, MF Has Smaller FPR than RF. MFI Plots Do Not Indicate Any Variable With Unusual High Score. 50

3.11 Null Scenario with Matching Associated with Exposure: (a) Comparison Between the Performance of MF, RF, and CLR in Terms of FPR, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated With $N = 600$ and $R = 100$. MF and CLR Have Small and Similar FPR and If $\alpha < 0.3$, MF Has Smaller FPR Than RF. MFI Plots Do Not Indicate Any Variable With Unusual High Score. 51

3.12 Null Scenario with Independent Variables: Effect of Number of Pairs (N) and Number of Exposure Variables (R) on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$ and $R \in \{20, 100, 150\}$. FPR of MF Is Small at Different Values of α and Is Not Sensitive to the Range of Values That We Tested for N and R. 51

3.13 Null Scenario With Matching Associated With Exposure: Effect of Number of Pairs (N) and Number of Exposure Variables (R) on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$ and $R \in \{20, 100, 150\}$. FPR of MF Is Small at Different Values of α and Is Not Sensitive to the Range of Values That We Tested for N and R. . . . 52

3.14 Linear Exposure: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated with a Linear Effect of x_1, $N = 600$ and $R = 100$. MF and CLR Are Accurate in Detecting the Linear Effect of x_1 and Their Roc Curve Dominates RF. MFI Plots Show That x_1 Has Substantially Higher MFI Score than the Other Exposure Variables and Matching Variables Have Small and Almost Identical MFI Scores. . 53

3.15 Linear Exposure: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_1 for Instances With Negative Effect ($\mu_1(-)$) on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$, $R \in \{20, 100, 150\}$, and $\mu_1(-) \in \{-0.5, -0.75, -1\}$ Respectively. MF Selects Variable x_1 As Important in Different Ranges of Values That We Tested and Its TPR Is Always Equal to 1 at Any Given Value of FPR. 53

3.16 Non-linear Exposure: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated with Non-linear Exposure Effect of x_1, $N = 600$ and $R = 100$. MF Performs Better than RF and CLR in Identifying the Non-linear Effect of x_1. MFI Plot Shows That x_1 Has Substantially Higher MFI Score than the Other Exposure Variables and All Matching Variables Have Small and Similar MFI Scores. 54

3.17 Non-linear Exposure: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_1 for Instances with Positive and Negative Effects $(\mu_1(+), \mu_1(-))$ on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$, $R \in \{20, 100, 150\}$, $(\mu_1(+), \mu_1(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$ Respectively. The Performance of MF Improves with Larger Number of Pairs, Effect Size and Smaller Number of Variables. 54

3.19 Exposure-exposure Interaction: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables for Data Sets Simulated with Exposure-exposure Interaction Between x_1 and x_2, $n = 800$ and $r = 100$. MF Performs Better than RF and CLR in Identifying the Correct Effect. MFI Plots Show That Both x_1 and x_2 Have Considerably Higher MFI Scores than the Other Exposure Variables and There Is No Matching Variable with Significantly Higher MFI Score than Others. 56

3.20 Exposure-exposure Interaction: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_2 for Instances with Positive and Negative Effects $(\mu_2(+), \mu_2(-))$ on MF Performance in Identifying An Exposure-exposure Interaction. Results Are Shown for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_2(+), \mu_2(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF Is Almost Accurate in All the Settings That We Tested for N and R and Its Performance Improves for Larger Effect Size. 57

3.21 Exposure-exposure-exposure Interaction: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables for Data Sets Simulated With Exposure-exposure-exposure Interaction Between x_1, x_2 and x_3, $N = 800$ and $R = 100$. The Performance of MF Is Better Than RF and CLR in Identifying the Correct Effects. MFI Plots Show That Variables x_1, x_2 and x_3 Have Received Considerably Higher MFI Scores Than the Other Exposure Variables and MFI Scores for Matching Variables Do Not Indicate Any Variable With Significantly Larger Score Than Others. All the Settings That We Tested for N and R and Its Performance Improves for Larger Effect Size. 58

3.22 Exposure-exposure-exposure Interaction: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_3 for Instances with Positive and Negative Effects $(\mu_3(+), \mu_3(-))$ on MF Performance in Identifying An Exposure-exposure-exposure Interaction Effect. Results Are Shown for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_3(+), \mu_3(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF Performs Well for All Values of N and R That We Tested and Its Performance Improves as Effect Size Increases. 59

Figure Page

4.1 Null Scenario Where Matching Variable v_1 Is Associated with Exposure Variable x_1: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iteration $t \in \{1, 2, 3\}$. MFI Scores of Exposure and Matching Variables Either Decrease Slightly or Remain Consistent with Increasing t. 73

4.2 Null Scenario Where Matching Variable v_1 Is Associated with Exposure Variable x_1: Evaluating the Performance of WMF in Terms of Its FPR along Different Values of Significance Level α. FPR of WMF Remains Consistent by Increasing t. 74

4.3 Null Scenario With Independent Variables: MFI Scores of (a) Exposure Variables And (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. MFI Scores of Exposure and Matching Variables Either Decrease Slightly or Remain Consistent with Increasing t. 74

4.4 Null Scenario with Independent Variables: Evaluating the Performance of WMF in Terms of Its FPR along Different Values of Significance Level α. FPR of WMF Remains Consistent by Increasing t. 75

4.5 Linear Exposure Effect: MFI Scores of (a) Exposure Variables And (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with a Linear Effect of x_1 Where $\mu_1 = -1$ for Pairs with Negative Effect Of x_1. MFI Score of Exposure Variable x_1 Is Larger than Other Exposure Variables in All Iterations and It Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variables Have Small MFI Scores Which Drop with Increasing t. 76

4.6 Linear Exposure Effect: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Is 1 for All Three Effect Size in Different Iterations. 77

4.7 Non-linear Exposure Effect: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with a Non-linear Effect of x_1 Where $\mu_1 \in \{-1, 2\}$. MFI Score of Exposure Variable x_1 Is Larger than Other Exposure Variables in All Iterations and It Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variables Have Small MFI Scores Which Drop with Increasing t. 78

4.8 Non-linear Exposure Effect: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves or Remains Consistent with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small. 79

4.9 Interaction Effect Between a Matching and An Exposure: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with An Interaction Effect Between Matching Variable v_1 and Exposure Variable x_1 Where $\mu_1 \in \{-1, 2\}$. MFI Score of Exposure Variable x_1 Is Larger than Other Exposure Variables in All Iterations and It Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variable v_1 Has Larger MFI Score than Other Matching Variables and It Improves with Increasing t, While the MFI Scores of Other Matching Variables Remain Consistent with Increasing t. 80

4.10 Interaction Effect Between a Matching and An Exposure: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small. 81

4.11 Interaction Effect Between Two Exposure Variables: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with An Interaction Between Exposure Variables x_1 and x_2 Where $\mu_1 = \mu_2 \in \{-1, 2\}$. MFI Score of Exposure Variables x_1 and x_2 Are Larger than Other Exposure Variables in All Iterations and Their Scores Improve with Increasing t, While the MFI Scores of Other Exposure and Matching Variables Remain Consistent or Decrease with Increasing t. 82

Figure Page

4.12 Interaction Effect Between Two Exposure Variables: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small. 83

4.13 Interaction Effect Between Three Exposure Variables: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with An Interaction Between Exposure Variables x_1, x_2 and x_3 Where $\mu_1(-) = \mu_2(-) = \mu_3(-) = -1$ and $\mu_1(+) = \mu_2(+) = \mu_3(+) = 2$. MFI Score of Exposure Variables x_1, x_2 and x_3 Are Larger than Other Exposure Variables in All Iterations and Their Score Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variables Have Small MFI Scores Which Drop with Increasing t. 84

4.14 Interaction Effect Between Three Exposure Variables: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small. 85

4.15 Effect Size Measures Based on Metrics M_1 and M_2 for a Variable Simulated with Negative Linear Effect. The Effect Size Is Estimated for $d \in \{-1.5, -1, -0.5, .5, 1, 1.5\}$. The Measure of Effect Size Based on M_1 Is Shown in Red (Triangles) on the Right Y-axis and the Measure Based on M_2 Is Shown in Blue (Circle) on the Left Y-axis. 88

Figure Page

4.16 Statlog Heart Disease Data Set: MFI Scores of WMF at $t \in \{1, 2, 3\}$. Variables x_3, x_4 and x_6 Received Relatively Larger Scores than Other Exposure Variables and Their MFI Scores Improve as t Increases. 90

4.17 Statlog Heart Disease Data Set: P-values Computed by WMF at $t \in \{1, 2, 3\}$. Variables x_5 and x_6 Were Selected by WMF in $t \in \{1, 2, 3\}$ at Significance Level $\alpha = 0.05$. 91

4.18 Childhood Acute Lymphoblastic Leukemia Study: MFI Scores of Top 50 Important Variable from the Third Iteration of WMF. 92

5.1 Linear Effect: (a) MNI Scores (Shap Values) of Exposure Variables from Matched Neural Network. (b) MFI Sores of Exposure Variables from Matched Forest. (c) SHAP Values of Exposure Variables from Conditional Neural Network. (d) Summary Plot of SHAP Values Estimated by Matched Neural Network. Each Point Corresponds to a Variable and an Instance, and the Color Represents the Feature Values from Low (Blue) to High (Red). (e) Effect of l and Effect Size ($|\mu_1^l|$) on the Accuracy of Matched Neural Network. 108

5.2 Interaction between Matching Variable v_1 and Exposure Variable x_1: (a) MNI Scores (SHAP Values) of Exposure Variables from Matched Neural Network. (b) MNI Scores (SHAP Values) of Matching Variables from Matched Neural Network.(c) MFI Sores of Exposure Variables from Matched Forest. (d) MFI Scores of Matching Variables from Matched Forest.e SHAP Values of Exposure Variables from Conditional Neural Network. (f) SHAP Values of Matching Variables from Conditional Neural Network..... 114

Figure Page

5.3 Interaction between Matching Variable v_1 and Exposure Variable x_1: (a) SHAP Dependence Plot to Show the Effect of d_1^{*1} Across All Data and Its Interaction with Matching Variable v_1 (b) Effect of l and Effect Size ($|\mu_1^l|$) on the Accuracy of Matched Neural Network. 115

5.4 Interaction between Exposure Variables x_1 and x_2: (a) MNI Scores (SHAP Values) of Exposure Variables from Matched Neural Network. (b) MFI Scores of Exposure Variables from Matched Forest. (c) SHAP Values of Exposure Variables from Conditional Neural Network. (d) Summary Plot of Shap Values Estimated by Matched Neural Network. Each Point Corresponds to a Variable and an Instance, and the Color Represents the Feature Values from Low (Blue) to High (Red). (e) Effect of L and Effect Size ($|\mu_1^l|$) on the Accuracy of Matched Neural Network. 116

5.5 Childhood Acute Lymphoblastic Leukemia Study: (a) The Median of MNI Scores (SHAP Values) over the 10 Replicates of Matched Neural Network for the Top 50 Genes with Largest Importance Scores. (b) The Impact of Variable $d^*_{208511\_at}$ Versus Its Value Across All Data. (c) The Impact of Variable 217099_s_at Versus Its Value Across All Data. . 117

6.1 The Scatter Plot of Control Versus Case for Exposure Variable x with No Effect in Original (a) and Transformed Data Set (b). 125

6.2 The Scatter Plot of Control Versus Case for Exposure Variable x with Linear Effect in Original (a) and Transformed Data Set (b). 126

6.3 The Scatter Plot of Exposure Variable x Simulated with No Effect for All Strata in Original (a) and Transformed Data Set (b). 130

Figure Page

6.4 The Scatter Plot of Exposure Variable x Simulated with an Effect for All Strata in Original (a) and Transformed Data Set (b). 131

6.5 Simple Scenario with One Important Exposure Variable x_1: Matched Importance Score of the Top 20 Exposure Variables with Largest Scores for Effect Size ($|\mu_1^0|$) $\in \{1, 1.5, 2\}$ and $T \in \{3, 5, 7\}$. 136

6.6 Complex Scenario with an Interaction Effect Between Two Exposure Variables x_1 and x_2: Matched Importance Scores of the Top 20 Exposure Variables with Largest Scores for Effect Size ($|\mu_2^0|$) $\in \{1, 1.5, 2\}$ and $T \in \{3, 5, 7\}$. 138

6.7 EEG Data Set: Matched Importance Scores of Exposure (a) and Matching (b) Variables. 140

Chapter 1

INTRODUCTION

In many applications, observations in data sets often show a dependency or grouping in their structure, where observations within a group or stratum tend to be correlated. Examples include blood pressure measured for each patient at different time points, measurements taken from different locations on each wafer, and subjects with and without diabetes matched together based on their age and gender. The valid analysis of these data sets requires methods that handle dependency structure among observations.

The focus of this dissertation is the analysis of matched study designs where the researcher is interested in identifying variables associated with a condition of interest and assess their effect. We use the terms outcome and condition of interest interchangeably in this dissertation. The outcome is nominal and units are sampled from populations corresponding to each outcome level. The variables which we would like to analyze their effect are referred as exposure variables. A special type of matched study designs is matched case-control study where observations are taken from a binary outcome including case or control. In clinical applications, case typically corresponds to units with a disease and control corresponds to those without the disease. In matched study designs, units are matched based on some baseline characteristics which are also referred as matching variables. That is, units within a stratum differ with respect to their outcome level but similar with respect to the values of matching variables.

As an example, consider a study that aims to identify the effect of diet on hearth disease using a matched study design where subjects with and without hearth disease

are matched based on their age and sex. Here, the outcome is binary; each unit is either with hearth disease or without hearth disease. The exposure variable of interest is diet and matching variables are age and sex.

This dissertation focuses on high-dimensional matched study designs where hundreds or thousands of exposure and dozens of matching variables with potential interaction among them exist. The interaction might be between multiple exposure variables or between matching and exposure variables. It is more appropriate to use the term effect modification for an interaction between matching and exposure variables, but to make it simple, we use the term interaction. For example, the childhood acute lymphoblastic leukemia study (Bhojwani *et al.* (2006)) uses a matched study design with 35 strata to evaluate 22283 genes (exposure variables) which are differentially expressed between the two states of the disease (diagnosis and relapse).

Matching is commonly used in observational studies to create a balance in the distribution of baseline variables in different outcome levels (Rothman *et al.* (2008)). Matching is performed on confounding variables which are associated with both the condition of interest and exposure variables (Rose and Van der Laan (2009)). If matching variable is not associated with either outcome and exposure variable, the efficiency of parameter estimations will be affected (Rose and Van der Laan (2009)). Also, matching variable should not be in the causal pathway between exposure and outcome, otherwise, the estimated parameters will be biased (Stuart (2010) and Rose and Van der Laan (2009)). The ideal method for matching is the exact matching method (Stuart (2010)) which creates strata such that the vector of matching values for units within a stratum have exactly zero distance. However, finding exact matches is difficult when number of matching variables is large. The common method for matching is the nearest neighbour method (Stuart (2010)). In the nearest neighbor matching, units with minimum distance between their matching vectors are selected

to create matched sets or strata. The purpose of this dissertation is not studying matching methods. Throughout this dissertation, we assume that matched data sets are provided and the exact matching method is used to create each stratum.

The analysis of matched data sets requires specific methods that account for matching structure of data. Paired t-test is the traditional method for analyzing matched data sets with binary outcome. It tests if the mean of an exposure variable differs between two outcome levels. This method is a univariate analysis which considers each exposure variable individually. It also does not test the existence of interactions among neither exposure variables nor matching and exposure variables. Tan *et al.* (2007) uses a modified version of paired t-test for a matched pair data set. We do not compare this with our proposed methods in this dissertation because of the limitations of paired t-test mentioned above.

When an outcome has more than two levels, the analysis of matched study designs is similar to the analysis of variance method for randomized block designs. Matched data sets have a similar structure as randomized block designs. The matching variables have the same role of blocking variables, and each stratum in matched study design can be considered as a block. In randomized block designs, the objective is to test if the mean of a variable differ among treatment levels. The objective of matched studies is also similar; we are interested in identifying if an exposure variable differ among different outcome levels. The analysis of variance method has similar limitations of paired t-test. This method cannot be used to test the existence of interaction between matching and exposure variables. Thus, we do not use it in this dissertation for comparison with our proposed methods.

The predominant method in the literature to analyze matched data sets with nominal outcome is Conditional Logistic Regression (CLR) model (Hosmer and Lemeshow (2000), Vierkant *et al.* (1999), Le Hesran *et al.* (2004) and Peleg *et al.* (2007)). CLR

accounts for the matched structure of data sets using a conditional likelihood approach which was introduced by Cox and Snell (1989). CLR is a multivariate linear model which estimates the effect of each exposure variable.

CLR has some disadvantages which limit its use in the analysis of matched data sets. First, to assess interaction effects between exposure variables or between exposure and matching variables, interaction terms (products of two or more variables) need to be included in the model. This increases the dimensionality of matched data sets significantly, especially when the number of variables is large. Second, it does not inherently handle categorical variables. We need to first convert categorical variables to one-hot encoded vectors, which increases the number of variables and leads to convergence problem. Other versions of CLR also have been proposed by Balasubramanian *et al.* (2014), Qian *et al.* (2014) and Asafu-Adjei *et al.* (2017) to handle high-dimensionality in matched data sets. These methods are linear and they still need interaction terms to capture their effects. When data set is high-dimensional, the number of added terms can be extremely large.

Throughout this dissertation, we propose methods to address the challenges exist in analyzing high-dimensional matched data sets. The methods are designed for the task of variable selection, effect estimation and classification in high dimensional matched data sets. Existing methods, as mentioned above, have difficulty in identifying interaction effects in high-dimensional data sets and they may not converge if the number of variables is excessive. Our proposed methods are able to inherently detect interaction effects without the need for adding the interaction terms. Thus, our models are more efficient for analyzing high-dimensional matched data sets.

In Chapter 3, we propose Matched Forest (MF) to address the problem of high-dimensionality in variable selection from matched case-control data sets. The method is designed for data sets with binary outcome where each stratum consists of one case

and one control. MF is a supervised machine learning algorithm based on Random Forest (RF) which inherently handles high-dimensionality without the need for including interaction terms in model. Also, it detects complex non-linear and interaction effects and handles both numerical and categorical variables. We demonstrate the effectiveness of MF in variable selection through extensive simulations and case studies. This work is published in Shomal Zadeh *et al.* (2020).

In Chapter 4, we propose three enhancements of Matched Forest (MF). First, we propose Weighted Matched Forest (WMF) to improve the variable selection accuracy in extremely high-dimensional data sets. WMF adaptively regularizes MF to focus on highly important variables. We show in our simulations and case studies that WMF outperforms MF in selecting important variables. Second, we generalize the application of MF to classification problems. we explain how MF can be used to classify instances in a matched pair to either case or control given the assumption that there is only one case and one control within each pair. Thus, the proposed algorithm classifies a pair as either case-control or control-case. Our experiments show that MF not only performs well in variable selection, but also has a better classification accuracy than competing algorithms. Finally, after important variables are identified, we explain how classification probabilities estimated by MF can be used to assess the effect of important exposure variables.

In Chapter 5, we propose Matched Neural Network (MNN) to assign importance scores to variables in high-dimensional matched case-control study designs. This method is suitable for both matched pairs with one case and one control within each stratum and matched $1 - L$ study designs where each stratum consists of one case unit and $L \geq 2$ control units. This method handles interactions inherently without any need to include interactions terms. We compared the performance of MNN with alternative methods in the literature including MF and observed in our

simulations and case studies that MNN performs better than alternative methods in identifying complex interaction effects. Another advantage of MNN compared to alternative methods is the use of interpretable SHAP scores (Lundberg and Lee (2017)) to measure the importance of each matching and exposure variable in matched data sets.

Chapters 3, 4 and 5 all propose methods for matched data sets with binary outcome. Chapter 6 focuses on matched data sets where outcome has more than two levels and proposes a machine learning model to assign importance scores to each matching and exposure variable in high-dimensional setting. Our empirical studies and analysis on a real data set show the advantages of our method in identifying important variables compared with existing methods in the literature.

Chapter 2

BACKGROUND

2.1 Matched Case-Control Study Design

Case-control study designs are commonly used in a wide range of applications as illustrated in Chapter 1 to identify exposure variables associated with a condition of interest. Case and control correspond to units with and without the condition of interest respectively. As an example in clinical application, consider a case-control study that aims to identify the effect of drinking coffee on hearth disease. Here, case group includes subjects with hearth disease present and control group includes subjects without hearth disease. The exposure variable of interest is drinking coffee. To estimate the effect of coffee on hearth disease, the amount of coffee consumed by subjects in these two groups is compared. Case-control studies are prone to selection bias which occurs when control samples are not representative of the population that produces the cases. To have a reliable comparison, case and control groups should be selected from a same population and only differ in the outcome of interest when there is no association between the disease and exposure variables. Matched case-control study designs are used to reduce this source of bias at the design stage.

A matched case-control study design is a special type of case-control studies where case and control subjects are grouped based on some matching variables. In particular, case subjects are matched with control subjects such that each group or matched set includes both cases and controls and subjects in a group have same values for the matching variables. The matched sampling increases the efficiency by making a balance in the distribution of confounding variables for cases and controls (Rothman

et al. (2008)). For example, in the aforementioned study, assume that smoking is a confounding variable which has different distribution in case and control samples. To remove the effect of smoking from analysis, matching is done to create homogeneous groups which are either smokers or non-smokers.

The number of case and control subjects within each group or stratum can vary, however, the most common designs include one case and 1 to 5 controls (Hosmer and Lemeshow (2000)). These data sets are called $1-L$ matched case-control study designs where L is the number of controls in each stratum. When each stratum includes one case and one control, it is also called 1-1 matched design and matched pairs. The application of this study design is widespread. For example, Balasubramanian *et al.* (2014) conducted a 1-1 matched study design to identify biomarkers associated with cardio-vascular disease after matching controls to cases on age, gender, race, ethnicity and severity of coronary artery disease.

2.2 Variable Selection Methods For Matched Case-Control Data Sets

2.2.1 Conditional Logistic Regression and Variants for High-dimensional Setting

The traditional approach to analyze matched case-control data sets is conditional logistic regression (CLR) which is a specialized type of logistic regression model (LR). CLR estimates the coefficients of each exposure variable based on a conditional likelihood function, but it cannot estimate the coefficients of matching variables. We first explain CLR for 1-1 matched design with one case and one control in each stratum. Then, we explain how it can be generalized to $1 - L$ matched designs with one case and $L > 1$ controls.

Consider a matched case-control study design with N matched pairs, R exposure variables, and M matching variables. Let $x^1(i)$ and $x^0(i)$ denote $R-$dimensional

vectors of exposure values for case and control subjects respectively and $V(i)$ denote an $M-$dimensional vector of matching values corresponding to pair i. Additionally, let $y^1(i)$ and $y^0(i)$ denote the case-control status of case and control subjects in pair i respectively, such that $y^j(i)$ ($j \in \{0,1\}$, $i \in \{1,2,\ldots,N\}$) takes 1 for cases and 0 for controls. The conditional likelihood for pair i is defined as

$$l_i(\beta) = \frac{e^{\beta' x^1(i)}}{e^{\beta' x^1(i)} + e^{\beta' x^0(i)}} \tag{2.1}$$

where $\beta' = \{\beta_1, \beta_2, \ldots, \beta_R\}$ is the vector of coefficients for R exposure variables. The coefficients of exposure variables (β) have the same interpretation as logistic regression (LR). The interpretation of each β_i is the change in logit for one unit increase in the corresponding exposure variable given all variables are constant within each matched pair. The conditional likelihood in equation 2.1 is in fact the probability that the case subject in pair i is actually a case given exposure values and under the assumption that one of the subjects in the pair is case. Based on this definition, $l_i(\beta)$ can also be represented as

$$l_i(\beta) = p(y^1(i) = 1 | y^1(i) + y^0(i) = 1, x^1(i), x^0(i)) \tag{2.2}$$

As can be seen in equation 2.1, the conditional likelihood function does not include any term corresponding to matching variables (V), thus, CLR cannot estimate the effect of matching variables and their effect can be estimated only when their interaction with other exposure variables is evaluated. The full conditional likelihood function is the product of $l_i(\beta)$ over N matched case-control pairs, namely,

$$l(\beta) = \prod_{i=1}^{N} l_i(\beta) \tag{2.3}$$

The coefficients β can be estimated by maximizing the conditional likelihood function in equation 2.3.

The conditional likelihood in equation 2.1 can be simply generalized for $1-L$ matched case-control study designs when $L > 1$ controls are matched with one case. To simplify the notation, consider a study where $L = 3$. We use $x^1(i)$ to denote exposure values of the case in stratum i and ($x^{01}(i)$, $x^{02}(i)$, $x^{03}(i)$) to denote exposure values of three control subjects in stratum i. The contribution of this stratum to the conditional likelihood function is obtained by

$$l_i(\beta) = \frac{e^{\beta' x^1(i)}}{e^{\beta' x^1(i)} + e^{\beta' x^{01}(i)} + e^{\beta' x^{02}(i)} + e^{\beta' x^{03}(i)}} \tag{2.4}$$

The interpretation of coefficients B is similar to equation 2.1 and the conditional likelihood function (l_i) computes the conditional probability that the subject with exposure values $x^1(i)$ is a case given the other three subjects ($x^{01}(i)$,$x^{02}(i)$, and $x^{03}(i)$) are controls.

CLR model is not suitable for high-dimensional matched case-control data sets with hundreds and thousands of exposure and matching variables. If we are also interested in non-linear and interaction effects, the dimensionality of data sets becomes even larger because we need to include non-linear and interaction terms (e.g. products of two or more variables and non-linear transformation of a variable) in data sets. CLR becomes intractable in such a high-dimensional setting. So far, variants of CLR have been introduced to handle high-dimensionality in matched case-control data sets and are presented as follows.

Random Penalized Conditional Logistic Regression (RPCLR): Balasubramanian *et al.* (2014) proposed an ensemble approach to assess variable importance in high-dimensional matched case-control data sets. To account for the matched structure of data, they use a penalized conditional logistic regression model. The method is proposed for matched pairs, but it can be extended to matched case-control $1-L$ designs where L is above 1. Specifically, their method consists of three major steps:

(i) a bootstrap sample is selected from matched pairs; (ii) for each bootstrap sample, a CLR model with ridge penalty is fitted to a random sample of variables with their pairwise interactions, and variable importance score of each variable is assessed; and (iii) finally, the importance score of each variable is computed as the average of scores over the bootstrap samples.

Penalized Conditional and Unconditional Logistic Regression: Qian *et al.* (2014) proposed a two-stage procedure which selects important variables from high-dimensional data sets in the first stage, and predicts outcome for future subjects in the second stage. For the variable selection method in stage one, first, main effects are identified by fitting a CLR model and then pairwise interaction of selected main effects is investigated by a penalized CLR model. For prediction in the second stage, the estimated coefficients for variables in stage one are used with unconditional logistic regression model to matched case-control data set. However, this method may fail to identify variables which do not have main effects but have pure interaction effects.

Bayesian Variable Selection Conditional Logistic Regression (BVS CLR): Asafu-Adjei *et al.* (2017) proposed a variable selection method for high dimensional matched case-control data sets which combines the advantages of CLR and Bayesian approaches. This method estimates the coefficients of variables and gives the probability estimates for inclusion of variables in the model, which can be used to rank variables. This method ignores interactions among variables including both matching and exposure. Although this method focuses on matched pairs design, Asafu-Adjei *et al.* (2017) claimed that it can be applied to more general $1 : L$ matched designs too.

2.2.2 Generalized Linear Mixed Models

An alternative method traditionally used for matched case-control studies is Generalized Linear Mixed Models (GLMM) (McCulloch and Neuhaus (2005)). For example, Szyszkowicz (2006) and Keogh (2017) applied GLMM on matched case-control data sets. This method uses a regression model which incorporates fixed effects for exposure variables and random effects for matching variables. Inclusion of random effects enables the heterogeneity of matched pairs in the analysis. GLMM also has similar disadvantages of CLR. It is not suitable for high-dimensional settings (large number of matching and exposure variables) where interaction among variables is also important. Interaction terms (cross products two or more variables) are needed to capture their effects. This increases the input dimension and can make this model intractable. Also, it does not inherently handle categorical variables. A conversion to binary variables increases the dimensionality further, especially with cross-product terms.

2.2.3 Boosting Methods

Adewale *et al.* (2010) proposed two variants of a boosting algorithm to classify matched pairs and select important variables from high-dimensional matched case-control data sets. The first method, Weighted L_2 Boosting ("WL_2Boosting"), combines gradient decent boosting algorithm with weighted L_2 loss function. The correlation between observations within each matched pair is handled through the matrix of weights which represents the unknown variance-covariance matrix of observations. The structure of this variance-covariance matrix is pre-specified and its parameters are estimated through an iterative procedure by minimizing the weighted L_2 loss function.

The second method, named 1-Step Penalized Quasi-Likelihood ("1-Step PQL-Boost") modifies the likelihood-based boosting algorithm (Tutz and Binder (2006)) via a generalized linear mixed model to handle correlation among observations in matched pairs. This method is similar to penalized quasi-likelihood (Breslow and Clayton (1993) and Molenberghs and Verbeke (2006)) that fits the linear mixed model on the pseudo data. However, instead of iterative fitting of linear mixed models as in PQL, the authors employ a one-step fitting. This boosting method also does not handle interactions among variables. After the classifier $F(X)$ is learned by each boosting method, the relative influence of each selected variable x_j is assessed using the following influence measure I_j proposed by Friedman (2001).

$$I_j = \left(E\left[\frac{\partial F(X)}{\partial x_j}\right]^2 .\mathrm{var}(x_j)\right)^{1/2} \tag{2.5}$$

The weak learner used in each iteration of both boosting algorithms is a simple linear regression model, and it does not handle interactions among variables.

2.2.4 Conditional Classification Algorithms

Stanfill *et al.* (2019) proposed a data transformation approach to generalize classification algorithms to matched case-control data sets. The new classification algorithms resulting from this data transformation are called conditional classification algorithms. The data transformation centers each strata by the mean values of exposures. For example, consider a case-control pair with one exposure variable which is measured as 750 and 250 for case and control respectively. After data transformation, the values of exposure will be 250 and -250 for case and control respectively. Formally, this approach is equivalent to the statistical practice of mapping a feature matrix into the null space. The authors employed this data transformation with 7 different linear and non-linear classification algorithms including Linear Discriminant

Analysis, CLR, Naive Bayes, Support Vector Machines with radial and linear kernels, Random Forest, and RPCLR. This method does not handle dependency among units within a stratum, which is recommended by statistical principles. It breaks each stratum into multiple instances which are known to be dependent. However, our proposed methods in this dissertation follow the statistical principles and construct data sets where instances are independent.

2.3 Variable Importance Measure Via Random Forest

Random Forest (RF) (Breiman (2001)) is an efficient and accurate machine learning algorithm for both classification and regression problems. RF is a combination of decision trees, each of which is grown on a bootstrap sample of training data set. Thus, each decision tree is built on a sample of data set, and the remaining data not used in a decision tree is called out-of-bag (oob) data. Another source of randomness in RF is injected when considering candidate variables for a split. In particular, instead of evaluating all variables for the best split, a random sample of variables is selected and then the best split is determined using this subset of variables. In R *randomForest* package, the default number of variables evaluated for the best split is $\sqrt{p}$ where p is the total number of variables in data set. The outcome for an unseen data is predicted by aggregating the predictions from all decision trees. For the classification problems, the final label is determined by majority votes and for regression problems, the final response is determined as the average of predictions over the decision trees.

RF is robust to overfitting as the number of decision trees increases (Breiman (2001)). The random sampling of variables at each node makes it an efficient algorithm when data set is high dimensional. Unlike many classification algorithms such as logistic regression, support vector machines and neural networks, RF inherently

handles both numerical and categorical variables and there is no need to convert categorical variables to binary one-hot encoded vectors before growing RF. Another interesting characteristic of RF is its ability to compute variable importance scores which can be used to select important variables.

Decision trees built by RF are based on the CART algorithm Breiman *et al.* (1984) which uses binary splits to divide the space. In classification setting, information gain based on the Gini index is used to evaluate candidate variables for the best split. Gini index at node ν is defined as

$$Gini(\nu) = 1 - \sum_{c=1}^{C} w_c^2$$

where C is the number of classes and w_c is the proportion of instances at node ν with class c. Gini information gain resulting from variable z at the parent node ν is defined as the difference between the Gini index at parent node (ν) and weighted average of Gini index of left (ν_l) and right (ν_r) child nodes. That is,

$$IG(z, \nu) = Gini(\nu) - p_l Gini(\nu_l) - p_r Gini(\nu_r)$$

where p_l and p_r are the proportions of samples in node v which are assigned to left and right nodes respectively.

RF has two measures for computing the variable importance score in a classification problem including Gini importance or Mean Decrease Gini (MDG) which is based on the information gain resulting from a split and Permutation importance or Mean Decrease Accuracy (MDA) which is based on the decrease in accuracy after permuting the values of a variable in oob data. In what follows, we explain the Gini importance in more details because this is used in this research. For more details regarding the Permutation importance of RF, see Breiman (2001) and Breiman (2002).

The R implementation of RF (Liaw and Wiener (2002a)) defines Gini importance of variable z ($VI(z)$) as the sum of weighted Gini information gain over nodes where variable z is used to split normalized by the number of trees. That is,

$$VI(z) = \frac{1}{ntree} \sum_{\{\nu | s(\nu)=z\}} N_\nu IG(z, \nu)$$

where $ntree$ is the number of decision trees in RF, $s(\nu)$ is the variable selected to split at node ν, and N_ν is the number of instances reaching node ν.

2.4 Variable Importance Measure Via Artificial Neural Networks

Neural network is a machine leaning algorithm which is extensively used in computer vision, natural language processing and time series prediction. There are different structures of neural networks including Multilayer Perceptron, Convolutional Neural Networks, and Recurrent Neural Networks, where each is suitable for a specific type of application. Figure shows the neural network structure for Multilayer perceptron (MLP). MLP consists of an input layer, an output layer, and at least one hidden layer with one or multiple neurons. This is a fully connected network because each neuron in a layer is connected to every neuron in the previous layer. Figure shows the neural network structure for MLP. Each neuron learns an output using a function of its inputs. Let $x_1, x_2, \ldots, x_H$ be inputs of neuron j, then its output o_j is computed as

$$o_j = f\left(\sum_{i=1}^{H} x_i w_{ij} + b\right) \tag{2.6}$$

where f is the activation function, w_{ij} is the weight of input x_i for neuron j, and b is the bias. Activation function can be as simple as a linear function or a more complex non-linear function such as rectified linear unit function (ReLU), Hyperbolic Tangent (TanH) and softmax. MLP learns the weights w_{ij} using backpropagation technique.

Neural networks achieve high accuracy in many applications and they can learn complex relationships between inputs and outputs. However, the black box nature of these models limits their use in applications such as healthcare and finance where interpretability is also important. A number of approaches have been proposed to address this challenge and to explain predictions of neural networks in terms of their inputs. These methods assign an attribution value, sometimes referred as contribution or relevance, to each input feature, which measures the contribution of that feature to the predicted output. In classification tasks, usually the output of interest is the one corresponding to the correct class. The feature attribution value computed by these methods is instance-level, and if all attributions are arranged in a two-dimensional matrix with the same size as input data, the resulting attribution will be called attribution map (Ancona *et al.* (2017)).

Shrikumar *et al.* (2017) and Ancona *et al.* (2017) provide an overview of feature attribution methods for neural networks. These methods are categorized as perturbation-based methods (Zeiler and Fergus (2014), Zhou and Troyanskaya (2015) and Zintgraf *et al.* (2017)), gradient-based methods (Simonyan *et al.* (2013), Sundararajan *et al.* (2016), Springenberg *et al.* (2014), Bach *et al.* (2015), Shrikumar *et al.* (2017) and Lundberg and Lee (2017)), and Grad-CAM and Guided CAM (Selvaraju *et al.* (2017)). Here, we explain one of the gradient-based methods, DeepSHAP Lundberg and Lee (2017), because it is used in Chapter 5 of this dissertation.

DeepSHAP is a feature attribution method designed for deep neural networks. It modifies DeepLIFT algorithm (Shrikumar *et al.* (2017) and Shrikumar *et al.* (2016)) to estimate SHAP (SHapley Additive exPlanations) values (Lundberg and Lee (2017)) over the feature space for each individual prediction.

SHAP is connected with Shapley values (Shapley (1953)) from game theory which explains the output of any machine learning model f by assigning an importance

value to each feature j (ϕ_j) that represents the effect of including that feature on prediction. To compute this effect for feature j, a model $f_x(S \cup j)$ is trained on a subset of features $S \subseteq F$, where F is the set of all features, with feature j present and another model $f_x(S)$ is trained on the feature subset S with feature j withheld. The marginal effect of feature j when it is added to feature subset S is then computed by $f_x(S \cup j) - f_x(S)$. When the model f is non-linear or feature variables are not independent, the marginal contribution of a variable depends on the other features in the model (S), thus, Shapley values arise from the weighted average of marginal contributions over all possible feature subsets $S \subseteq F \setminus \{j\}$:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup j) - f_x(S)] \tag{2.7}$$

For many of the machine learning models, it is not feasible to predict the output of the model for a subset of features. SHAP uses a conditional expectation function of the original model to define simplified input features. That is, it defines $f_x(S)$ by $E[f(x)|x_S]$ which is the expected value of the model conditioned on the feature subset S. The exact computation of SHAP values is challenging for complex models. However, existing additive feature attribution models can be modified to approximate SHAP values.

DeepLift is an additive feature attribution method for neural network which is modified in DeepSHAP to approximates SHAP values. DeepLift is one of backpropagation-based approaches that propagate an importance signal from an output neuron through hidden layers and finally to input features. This is computationally efficient because importance scores are computed in only one backward pass. Let y be a target neuron which is defined as $y = f(x)$ and x be a vector of n features $\{x_1, x_2, \ldots, x_n\}$. DeepLIFT assigns to each feature x_i a contribution score $C_{\Delta x_i \Delta y}$ that represents the amount of difference in output y from a reference attributed to

the difference of that feature from the reference. The choice of a reference depends on domain-specific knowledge. For example, Shrikumar *et al.* (2017) uses an image with all zeros as the reference for MNIST data set because this is the background of all images in this data. DeepLIFT is an additive feature attribution method that follows "summation-to-delta" property:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta y} = \Delta y \tag{2.8}$$

where $\Delta y = f(x) - f(x^0)$ and $\Delta x_i = x_i - x_i^0$. Similar to how chain rule is constructed for partial derivatives to compute the gradient of the output with respect to an input, DeepLIFT uses "chain rule for multipliers" to compute the global multiplier for any neuron to a given target neuron via backpropagation. For a given input x and target neuron y, the multiplier is defined as:

$$m_{\Delta x \Delta y} = \frac{C_{\Delta x \Delta y}}{\Delta x} \tag{2.9}$$

which is the contribution of input x to target neuron y divided by the difference-from-reference of the input Δx. When Δx is close to zero, the definition of multiplier will be similar to partial derivative $\frac{\partial y}{\partial x}$ which is the change in y caused by an infinitesimal change in x divided by the infinitesimal change in x. According to the chain rule for multipliers, the global multiplier from input x to target neuron y ($m_{\Delta x \Delta y}$) is computed by recursively passing the multipliers backward through the network and summing them up over all paths connecting input x to target neuron y. Assuming that there is a hidden layer with neurons $h_1, \ldots, h_n$ between input neuron x and target neuron y, the global multiplier for x to y is computed as follows:

$$m_{\Delta x \Delta y} = \sum_i m_{\Delta x \Delta h_i} \times m_{\Delta h_i \Delta y} \tag{2.10}$$

and the the contribution score of input neuron x to target neuron y is computed as

$$C_{\Delta x \Delta y} = m_{\Delta x \Delta y} \times \Delta x \tag{2.11}$$

Shrikumar *et al.* (2017) introduces some rules including linear, rescale and reveal cancel to compute the multiplier for each neuron to its immediate inputs. These rules are suitable for activations with linear functions or nonlinear functions with only one input. Non-linear functions with multiple inputs are not addresses in Shrikumar *et al.* (2017), and the public implementation of DeepLIFT uses the gradient for such functions (Ancona *et al.* (2017)).

DeepSHAP modifies DeepLift by computing SHAP values for smaller components of neural network analytically and propagate them backward through the network using DeepLift's multipliers. If we interpret the reference input in Equation 2.8 by $E[x]$ (expected value of inputs), DeepLift approximates SHAP values assuming that input feature are independent and neural network model is linear.

Chapter 3

MATCHED FOREST: SUPERVISED LEARNING FOR HIGH-DIMENSIONAL MATCHED CASE-CONTROL STUDIES

3.1 Introduction

Matched case-control designs are commonly used in a wide range of applications as illustrated in Chapter 1 to remove the effect of confounding variables in identifying important variables associated with a condition of interest. Case and control correspond to units with and without the condition of interest respectively. Case and control instances are grouped into a stratum based on some matching variables and a number of exposure variables are studied for their effect on the condition of interest. For example, Heller *et al.* (2008) aims to identify differentially expressed genes between two subgroups of leukemia patients after matching on variables including age, sex, multi-drug resistance (mdr), the stage of cell differentiation (stage) and an indicator variable of whether the chromosome number was large.

We should include all variables as matching which are associated with both exposure and health condition (case or control) (Rose and van der Laan (2009)). If a matching variables is only associated with either the health condition or exposure variables, the variance will increase (Rose and van der Laan (2009)). Also, we should not match on variables which are affected by the exposure to avoid post-treatment bias (Ho *et al.* (2007) and Rose and van der Laan (2009)).

Traditionally, matched case-control data sets were analyzed by Conditional Logistic Regression (CLR) (Hosmer and Lemeshow (2000)) which tests the significance of each exposure variable in the full logistic regression model including all variables. For

example, Vierkant *et al.* (1999), Le Hesran *et al.* (2004) and Peleg *et al.* (2007) all applied CLR to analyze matched case-control data, with applications to epidemiology (based on a conditional likelihood function Cox and Snell (1989)). An alternative approach is a Generalized Linear Mixed Model (GLMM). For example, Szyszkowicz (2006) and Keogh (2017) applied GLMM on matched case-control data sets. This method uses a regression model which incorporates fixed effects for exposure variables and random effects for matching variables. Inclusion of random effects enables the heterogeneity of matched pairs in the analysis. Both CLR and GLMM have some disadvantages which limit their use in matched case-control analysis. Difficulties occur with high dimensional data with a large number of exposure variables and interaction between variables is of interest. Interaction terms (cross products two or more variables) are needed to capture their effects. This increases the input dimension and can make these models intractable. CLR also runs into convergence problem for high dimensional data (Asafu-Adjei *et al.* (2017)). Also, both CLR and GLMM do not inherently handle categorical variables. A conversion to binary variables increases the dimensionality further, especially with cross-product terms.

High dimensionality was considered by Balasubramanian *et al.* (2014), Qian *et al.* (2014), Adewale *et al.* (2010), Tan *et al.* (2007) and Asafu-Adjei *et al.* (2017). These methods basically adapt linear models, sometimes supplemented with some cross-product terms to consider interactions. In high dimensional data, the added terms can become excessive. Our method is quite different, with an inherent capability to handle interactions. Also, the Random Forest (RF) algorithm (Breiman (2001)) is a popular method for variable selection in high dimensional data sets. However, RF does not take into account the matched structure of data. Balasubramanian *et al.* (2014) showed some limitations of RF in analyzing matched case-control data and we also illustrate weaknesses in experiments relative to our method.

We present a quite different approach for variable selection in high dimensional matched case-control studies.Our key idea is to transform data based on the potential outcome model (Neyman (1923), Rubin (1977)). A new label is defined and a supervised learner is applied to the transformed data with modified variable importance (VI) scores. We use a RF for our learner due to its ability to identify complex interaction effects in high dimensional settings and provide VI scores. The approach is conceptually simple and computationally scales well. Experiments demonstrate the effectiveness of the proposed approach in the presence of large number of matching and exposure variables.

Section 3.2 presents background on matched case-control analysis, existing methods for high dimensional settings and RF VI scores. Section 3.3 describes the method. Section 3.4 presents the results from experiments on simulated and biomedical data sets. Section 3.5 provides conclusions.

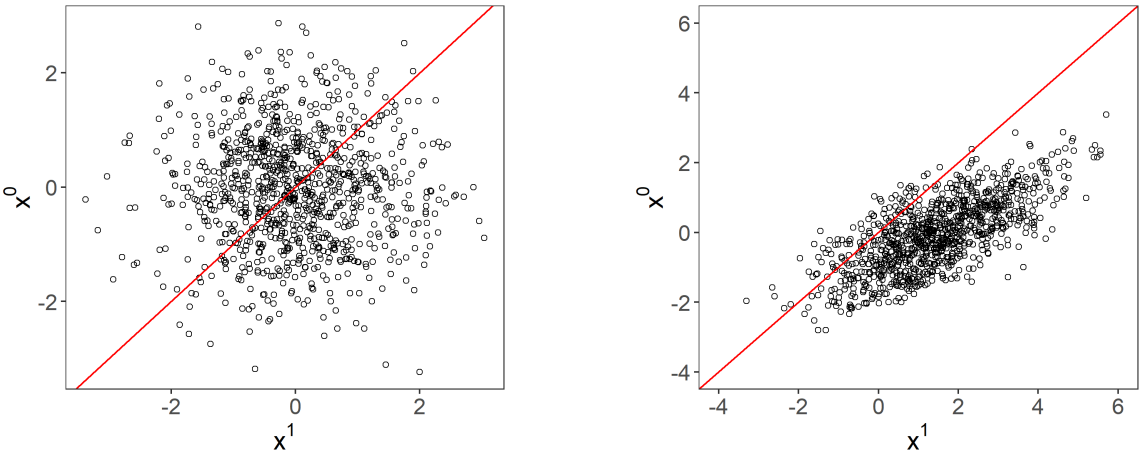
3.2 Background

3.2.1 Matched Case-Control Analysis

Consider data where subjects are paired based on a number of matching variables and exposure variables are to be studied for their effects on a binary outcome. As an example, subjects can be paired based on age, the exposure variable of interest is diet, and the subjects are evaluated for heart disease (outcome of interest). For every subject with heart disease present (a case), there is a subject selected of the same age without heart disease (a control). The analysis is to study the relationship between heart disease and diet.

Let x represent the average daily calorie intake and x^0 and x^1 denote the values for the control and the case, respectively. Figure 3.1a shows a scatter plot of x^0 versus

x^1 for each pair from a simulated data set. From the figure, there does not seem to be regions where x differs between the cases and the controls. That is, we do not detect regions in the x^0 and x^1 space where the value of case is consistently greater/smaller than the value of control. Hence, no effect of diet is observed in the plot.



(a) No effect

(b) Exposure effect

Figure 3.1: Scatter Plot of Control (x^0) Versus Case (x^1) for Exposure Variable x with (a) No Effect and (b) Effect

Now, consider the example from a simulated data set in Figure 3.1b. Many pairs are present with $x^1 > x^0$ that indicate an effect of diet. However, standard supervised learners that do not consider matching have difficulty identifying this effect because the methods compare the overall distribution of x^1 and x^0 and they do not consider how case and control values differ within each pair.

Sometimes the effect is more subtle. Consider the example in Figure 3.2 from a simulated data set with an interaction effect between an exposure and a matching variable (denoted as v and binary such as male or female patients). In Figure 3.2, blue (circles) and green (triangles) represent male and female groups, respectively. The majority of pairs in male group have x^1 smaller than x^0, while the majority

of pairs in female group present the opposite effect. Therefore, there is an effect from the exposure variable, however, it is interacted with the matching variable v. If the matching is not considered, we may incorrectly conclude that there is no effect because the number of pairs with case greater than control are almost equal to the number of pairs with control greater than the case for exposure variable x.

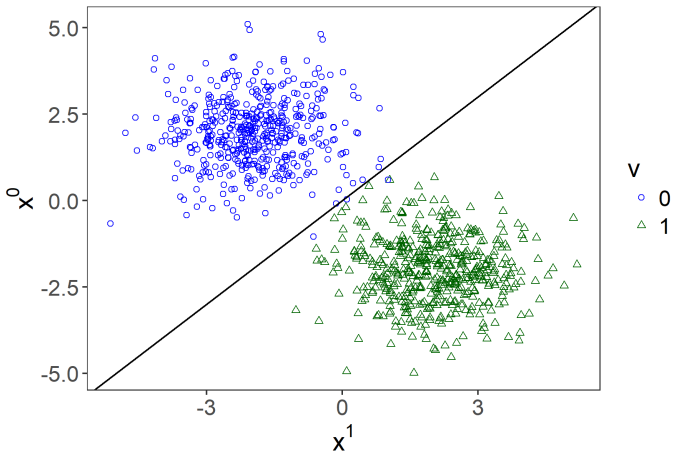


Figure 3.2: Scatter Plot of Control Versus Case for an Exposure Variable x with Matching-exposure Interaction Effect. Blue (Circles) and Green (Triangles) Represent Male (Encoded as 0) and Female (Encoded as 1) Groups of the Matching Variable, Respectively.

3.2.2 Existing Methods For High-Dimensional Matched Case-Control Data Sets

The problem of high dimensionality in matched case-control analysis has been considered by Tan *et al.* (2007), Adewale *et al.* (2010), Balasubramanian *et al.* (2014), Qian *et al.* (2014) and Asafu-Adjei *et al.* (2017). Tan *et al.* (2007) developed a two-stage variable selection approach, where in the first stage a modified t-test statistic is used, and in the second stage, a support vector classifier based on selected variables is built. However, the variable selection procedure uses a univariate analysis

which does not include interaction effects. Also, matching is not incorporated in the classification model. Adewale *et al.* (2010) proposed two variants of a boosting algorithms for matched case-control analysis. The first method combines a generic gradient boosting algorithm with a weighted least square loss function that handles correlated binary outcomes. The second method modifies a likelihood optimization boosting algorithm by using GLMMs. However, both boosting algorithms do not include interaction terms in their model. Balasubramanian *et al.* (2014) used an ensemble approach. For each bootstrap sample, a CLR model with ridge penalty is fit to a random subset of features with their pairwise interaction. This model only includes interaction terms between two variables, thus, it does not detect interaction effects which include larger number of exposure and matching variables. This method also had some convergence problems in our implementation for our experiments. Qian *et al.* (2014) developed two variable selection approaches based on conditional and unconditional logistic regression as well as lasso and ridge penalties. The first method employs a procedure which selects variables by CLR, and then uses unconditional logistic regression for prediction. For variable selection at stage one, first, important main effects are identified, and then pairwise interaction of selected variables are investigated. However, this method may not identify variables which do not have main effects, but more pure interaction effects, or higher order interactions (¿2) among exposure and matching variables. Their second method performs variable selection and prediction simultaneously by fitting an unconditional logistic regression model. Although the second model includes both matching and exposure variables with their interactions, it becomes intractable for high dimensional data sets because of the addition of many interaction terms. Asafu-Adjei *et al.* (2017) proposed a Bayesian variable selection approach for matched case-control analysis which is formulated in

a CLR framework. This method also ignores interaction effects between variables including both matching and exposure.

Breiman's Random forest (RF) (Breiman (2001)) is an efficient supervised method for assessing VI in high dimensional data sets. RF naturally handles different scales, non-linear effects, interactions, categorical and numerical variables. Additionally, RF computes VI score for each variable in data set. As mentioned, researchers used RF to analyze matched case-control data Tsou *et al.* (2007). However, they did not account for matching in their analysis.

RF consists of several trees each is built on a bootstrap sample of data set. A subset of variables are randomly selected and evaluated to split instances at each node. Most implementations of RF offer two measures for variable importance including Gini importance which is based on the decrease in the impurity after the split and the permutation importance which is based on the decrease in accuracy after permuting values of a variable (Strobl and Zeileis (2008)). As this paper uses the Gini importance measure, we explain this in more details. If variable z is selected to split N instances at node ν into left child node ν_l and right child node ν_r with proportions of instances $p_l = N_l/N$ and $p_r = N_r/N$, respectively, the Gini information gain of variable z for splitting at node ν ($IG(z,\nu)$) is computed as $IG(z,\nu) = Gini(\nu) - p_l Gini(\nu_l) - p_r Gini(\nu_r)$. Gini impurity at node ν (Gini(ν)) is defined as $1 - \sum_{c=1}^{C} w_c^2$ where C is the number of classes and w_c is the proportion of instances at node ν with class c. The R implementation of RF (Liaw and Wiener (2002b)) computes importance measure for variable z denoted as $VI(z)$ by the sum of weighted Gini information gain over all nodes where variable z is used to split normalized by the number of trees:

$$VI(z) = \frac{1}{ntree} \sum_{\nu \in \{\nu | s(\nu) = z\}} N_\nu IG(z,\nu)$$

where *ntree* is the number of trees in random forest, $s(\nu)$ is the variable selected to split at node ν, and N_ν is the number of instances reaching node ν.

3.3 Matched Forest

We introduce a new algorithm, Matched Forest (MF), for variable selection in high dimensional matched case-control data set. Existing models for matched case-control data require additional terms (cross products) to be explicitly added to handle interactions. But this increases the dimensionality significantly and make variable selection a more difficult problem. MF is both simple conceptually and easy to implement, yet based on the established potential outcome model (Rubin (1977)). MF consists of two simple steps: 1) a transformation to convert the variable selection problem into a supervised setting based on the potential outcome framework for causal inference; 2) a supervised learner which is able to inherently detect the complex interactions involving both exposure and matching variables in high dimensional setting using the transformed data set.

3.3.1 Transform to Supervised Learning

A task for causal effects is to use observed exposure values to estimate unobserved ones which are also known as counterfactuals. In matched case-control study designs, because subjects within each pair are similar to each other in terms of matching variables, we can use the control's observed exposure to represent the case's counterfactual (He *et al.* (2016)). Similarly, the case's observed exposure is used to represent control's counterfactual.

Formally, suppose that (i_1, i_0) denotes, respectively, case and control subjects in pair i. The observed values of exposure for case and control subjects in this pair are $x^1(i_1)$ and $x^0(i_0)$, respectively, and the counterfactuals are $x^0(i_1)$ and $x^1(i_0)$. Here

$x^0(i_1)$ is the potential exposure value if the case subject were the control (analogously $x^1(i_0)$). To estimate the individual causal effect for pair i which is defined as $\delta_i = x^1(i_1) - x^0(i_1)$, we need to first estimate the counterfactual $x^0(i_1)$. In matched case-control study designs, the control's observed exposure $x^0(i_0)$ is used to estimate cases's counterfactual $x^0(i_1)$. Thus, $\hat{\delta}_i = x^1(i_1) - x^0(i_0)$ is an estimate of individual causal effect. The average effect of exposure is then estimated from the mean difference in a sample of N matched case-control pairs as

$$\hat{\delta} = \frac{1}{N}\sum_{i=1}^{N}[x^1(i_1) - x^0(i_0)]$$

Our method compares the observed values of exposure and counterfactuals, but with a different approach that extends beyond the average of differences and can detect more complex relationships between exposures and outcomes.

Because we only need the observed exposure values to estimate the causal effect, we simplify notation to $x^1(i)$ and $x^0(i)$ to denote, respectively, the observed exposure values $x^1(i_1)$ and $x^0(i_0)$ for case and control subjects in pair i. Let N denote the number of matched case-control pairs. We generate new case and control variables, denoted as (x^{*1} and x^{*0}), with $2N$ rows. For the first N rows, the instances match the original instances and for the second set of N rows, case and control values of the exposure are interchanged within each pair. That is,

$$x^{*k}(i) = \begin{cases} x^k(i) & \text{for } i = 1, 2, \ldots, N, \\ x^{1-k}(i-N) & \text{for } i = N+1, N+2, \ldots, 2N \end{cases}$$

for $k = \{0, 1\}$. The second N rows represent the counterfactuals (estimated based on the potential outcome model for a matched case-control study design). Because the effect of an exposure is dependent on the difference between its case and control values within each pair, if the exposure variable is numerical, we also create new

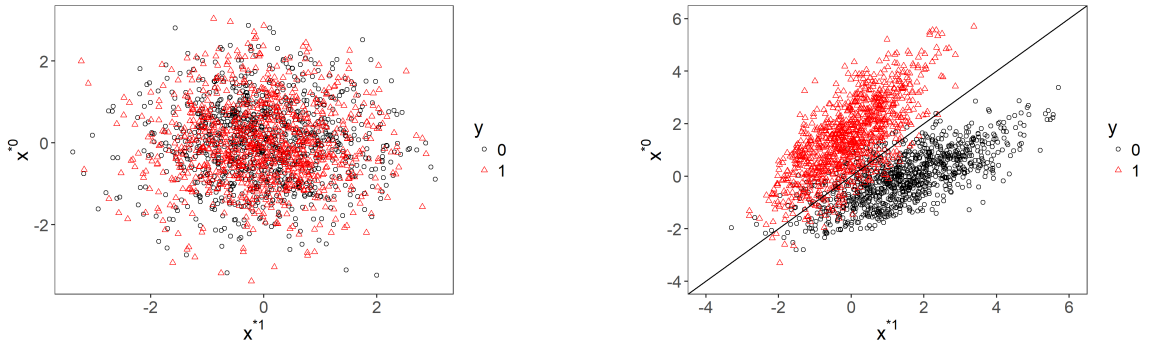
variable d^* defined as

$$d^*(i) = x^{*1}(i) - x^{*0}(i)$$

for $i = 1, 2, \ldots, 2N$. We distinguish the observed and counterfactual pairs with a label, defined as $y(i)$ equals 0 and 1 for the original and counterfactual rows, respectively. MF uses a classifier on the transformed data set consisting of x^{*1}, x^{*0}, d^* and y to evaluate the effect of exposure x. If the exposure variable x has an effect on outcome ($\delta \neq 0$), we would expect the classifier to separate the original and counterfactual pairs.

To illustrate our method, recall the examples in Figures 3.1a and 3.1b. In Figures 3.3a and 3.3b, we plot x^{*0} versus x^{*1} with observed exposure values in black (circles) and counterfactuals in red (triangles), respectively. In Figure 3.3a the observed and counterfactuals are not clearly separated so that a supervised learner with x^{*1}, x^{*0} and d^* would not classify well. In our approach, this indicates no effect for exposure variable x. However, in Figure 3.3b, many of the observed exposure values are below the 45 degree line while their counterfactuals are above this line. Thus, a supervised learner can use x^{*1}, x^{*0}, and d^* to classify well. Variable d^* is useful to separate observed and counterfactual pairs because its value is positive (negative) for many of the original (counterfactual) pairs. For a numerical predictor, potentially either x^{*0} or x^{*1} could be removed from the analysis when d^* is included, but we prefer to keep both to maintain the symmetry and a general approach.

Our method can be easily extended to a number of exposure and matching variables, with subsequent major advantages to handle multiple effects and detect interactions. Consider a data set with R exposure variables denoted by $x_1, x_2, \ldots, x_R$ and M matching variables denoted by $v_1, v_2, \ldots, v_M$. We denote the case and control values for an exposure variable x_r as x_r^1 and x_r^0, respectively. As described previously, our approach adds counterfactuals and variables to the original data set to generate



(a) No effect

(b) Exposure effect

Figure 3.3: Scatter Plot of New Control (x^{*0}) Versus New Case (x^{*1}) Variables Associated with Exposure Variable x with (a) No Effect and (B) An Effect. Observed Exposure Values ($y = 0$) and Counterfactuals ($y = 1$) Are Shown by Black (Circles) and Red (Triangles).

new variables, denoted by x_r^{*1} and x_r^{*0}. If our data set has R_1 numerical exposure variables, then we also generate difference variables

$$d_r^* = x_r^{*1} - x_r^{*0}$$

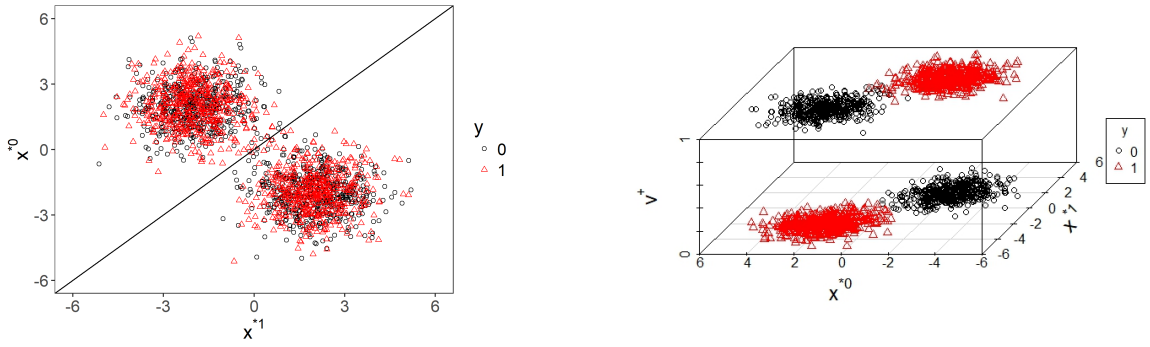
for $r = 1, 2, \ldots, R_1$. To detect interaction effects between matching and exposure variables, additional variables are generated. We create new matching variables v_m^+ for $m = 1, 2, \ldots, M$ by extending the original matching variables (v_m) to $2N$ instances as

$$v_m^+(i) = \begin{cases} v_m(i) & \text{for } i = 1, 2, \ldots, N, \\ v_m(i - N) & \text{for } i = N + 1, N + 2, \ldots, 2N \end{cases}$$

for $m = 1, 2, \ldots, M$. Therefore, the transformed data set has $2N$ instances and $M + 2R + R_1 + 1$ columns.

To illustrate the role of v_m^+, recall the example in Figure 3.2. In Figure 3.4a, we plot x^{*0} versus x^{*1} with the observed in black (circles) and counterfactuals in red (triangles). Although x has an effect, Figure 3.4a shows that the observed and

counterfactuals are not clearly separated However, when the matching variable v^+ is included in the data and we view the three-dimensional plot of v^+, x^{*1} and x^{*0} in Figure 3.4b, the observed and counterfactuals are separated and a classifier using these variables would perform well.



(a) No effect

(b) Exposure effect

Figure 3.4: Scatter Plot of Control x^{*0} Versus Case x^{*1} (a) Without and (b) with Matching Variable v Illustrate An Interaction Effect Between x and v. Observed Exposure Values ($y = 0$) and Counterfactuals ($y = 1$) Are Shown by Black (Circles) and Red (Triangles).

3.3.2 MF Variable Importance

Our MF algorithm applies RF on the transformed data set to identify important matching and exposure variables. One of the advantages of our method is that widely available algorithms can be simply applied. We selected RF due to its ability to handle high dimensionality, mixed variables, and interaction effects and to measure VI scores which can be used for variable selection.

To select important variables, we compute a MF variable importance score, denoted as MFI, for each variable based on the Gini method, but modified for the matched analysis. The MFI score for matching variable v_m for $m = \{1, 2, \ldots, M\}$ is

computed as $MFI(v_m) = VI(v_m^+)$. The MFI score for exposure variable x_r is computed by summing up the Gini importance scores of all case, control and difference variables related to exposure variable x_r as

$$MFI(x_r) = VI(x_r^{*0}) + VI(x_r^{*1}) + VI(d_r^*)$$

All these variables are derived from exposure variable x_r and can appear as splitting variables in trees. Therefore, the summation is used to provide an overall score for the importance of x_r.

The computational cost for MFI scores is the same as VI scores and it depends on the number of pairs (N), the number of variables selected at each split ($mtry$) and the number of trees ($ntree$). Thus, for a fixed $mtry$, MF and RF have the same computational cost for computing MFI and VI scores respectively. In the default parameter setting of R randomForest package, $mtry$ is set to ($\sqrt{p}$) where p is the total number of variables in data set. The data transformation step in MF increases p linearly in the number of matching and exposure variables which does not change the computational cost.

3.4 Experiments

We compared MF with either CLR, RF, or Boosting Weighted L_2 Loss (WL_2Boost) (Adewale *et al.* (2010)) in a series of experiments. In section 3.4.1, simulations are used to demonstrate the performance of MF. The full description of simulations and results are provided in the supplementary information. In section 3.4.2, four biomedical data sets are analyzed to evaluate MF.

3.4.1 Simulation Studies

We tested the performance of MF in variable selection through extensive simulation studies. In particular, we generated data sets with no effect of exposure and five different effect types including linear, non-linear, matching and exposure interaction, two exposures interaction and three exposures interaction. In our simulations, number of pairs (N) and number of exposure variables (R) range from 300 to 1000 and 20 to 150, respectively. For all simulation studies, unless otherwise stated, each control variable x_r^0 for $r = 1, 2, \ldots, R$ is generated randomly from a uniform distribution between 1 and 50 and each case variable x_r^1 for $r = 1, 2, \ldots, R$ is generated according to $x_r^0 + d_r$ where d_r follows $N(\mu_r, 1)$. In our simulations, μ_r is set to 0 for exposure variables without any effect, to $\{-0.5, -0.75, -1\}$ for exposure variables with negative effect and to $\{1, 1.5, 2\}$ for exposure variables with positive effect. Let the absolute value of μ_r ($|\mu_r|$) be the effect size. Larger values for $|\mu_r|$ indicate stronger effects of exposure variables. Number of instances with negative, positive and no effect are also changed for some variables to generate different types of effects. Variables in interaction effects are simulated so that they individually do not have an effect on the outcome, but their combination with other variables show an effect. Moreover, matching variables are generated independently from Poisson (5) distribution. This provides discrete matching values with common values that might be expected in practice (e.g., age intervals).

For each simulation study, we generated 100 data sets and compared the performance of variable selection from MF, RF, and CLR. To select important variables from MF (RF), we compared MFI (VI) scores to the null distribution generated from randomly assigned labels. In particular, for each simulated data set, we permuted the case and control instances within a pair for each exposure variable and computed

MFI (VI) scores on the new data sets to estimate the null distribution for each MFI (VI) score. Observed MFI (VI) scores using the original data set are compared with the estimated null distributions and variables with MFI (VI) score significantly large at level α are selected as important. We did not adjust α for the multiple comparison problem in our simulations, although the problem exists due to the large number of variables. The objective is to show the relative performance of MF to the alternatives without model-building for simulated data sets.

We used default parameters in R randomForest package to run MF. Specifically, we set $ntree = 500$, $mtry = \sqrt{p}$, and grow trees to purity. To evaluate the performance of variable selection, we used ROC curve which plots true positive rate (TPR) versus false positive rate (FPR) along different values of significance level α. Let I and I' be sets of important and noise variables (both exposure and matching) respectively and $N_\alpha(z)$ be the number of simulations out of 100 simulated data sets that selected variable z as important at significance level α. Then, TPR and FPR are defined respectively as

$$TPR = \frac{\sum_{z \in I} N_\alpha(z)}{100|I|} \quad \text{and} \quad FPR = \frac{\sum_{z \in I'} N_\alpha(z)}{100|I'|}$$

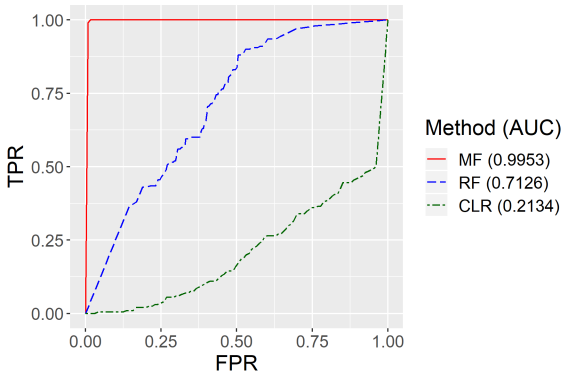
where $|\{.\}|$ denotes the cardinality of set $\{.\}$. In our simulations, α ranges from 0 to 1 incremented by 0.01. As CLR is not able to estimate the effect of matching variables, $N_\alpha(z) = 0$ for $z \in \{v_1, v_2, \ldots, v_M\}$. Our experiments show that MF outperforms CLR in detecting non-linear and interaction effects, however, they have similar performance when the exposure variable is linearly associated with the outcome. Also, MF performs better than RF in identifying different effect types. Here, we show results for the simulation study with an interaction between a matching and an exposure variable. For a full description of simulations and results, see the supplementary information.

We simulated an interaction effect between exposure variable x_1 and matching variable v_1 with no other effects. Values for v_1 are generated from a Poisson (5) distribution and sorted in an ascending order. Variable x_1^1 is generated to have positive effect for smaller values of v_1 and negative effect for higher values of v_1. That is, for the first $N/4$ instances, $\mu_1 = 2$ (positive effect), for the next $N/4$ of instances, $\mu_1 = 0$ (no effect) and for the last $N/2$ instances $\mu_1 = -1$ (negative effect).The mean difference is 0 ($\hat{\delta}_1 = 0$) so that x_1 individually does not have an effect. Also, $\mu_r = 0$ for $r = \{2, \ldots, R\}$. Figure 3.5a shows ROC curves from MF, RF, and CLR and Figures 3.5b and 3.5c show MFI scores of matching and exposure variables respectively for data sets with a matching-exposure interaction, $N = 800$ and $R = 100$. From Figure 3.5a, we can see that MF is more accurate in identifying important variables compared to RF and CLR because its ROC curve dominates the other two methods. Also from Figure 3.5b and Figure 3.5c, we observe that matching variable v_1 and exposure variable x_1 both have higher MFI scores than other exposure and matching variables.

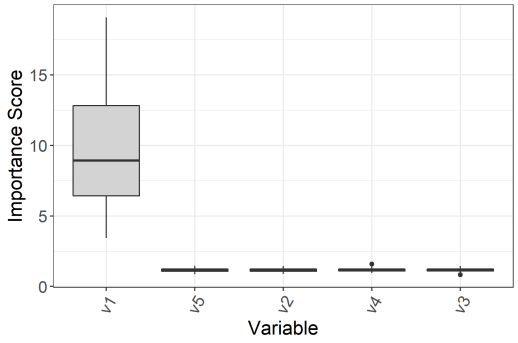
3.4.2 Biomedical Examples

Three unmatched data sets from UCI benchmark database Lichman (2013) and one paired gene expression data set were used for analysis. The unmatched data sets were converted to a matched design to be used for matched case-control analysis. Age and gender are two variables commonly used for matching in clinical studies. For each data set, we used at least one of these variables to match controls to similar cases. We used the R package *Matchit* for exact matching of controls to cases. Instances that were not matched were removed from the analysis. Further details regarding matching for each data set are described in the corresponding section.

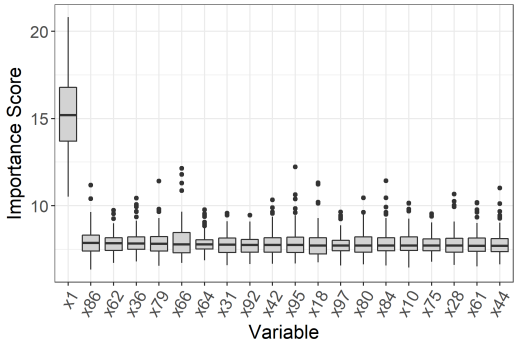
We ran MF on each data set 10 times to account for randomness of MF in computing the MFI scores. Similar to section 3.4.1, important variables are selected by



(a) ROC curve



(b) MFI scores: matching variables



(c) MFI scores: exposure variables

Figure 3.5: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables, and (c) MFI Scores of Exposure Variables for Data Sets Simulated with Matching-Exposure Interaction Between v_1 and x_1, $N = 800$ and $R = 100$. MF Performs Better than RF and CLR in Identifying the Correct Effect. MFI Plots Show That Both x_1 and v_1 Have Substantially Higher MFI Scores than the Other Exposure and Matching Variables.

comparing the average of MFI scores over 10 runs of MF with the estimated null distributions. We permuted instances in a pair 100 times to estimate the null distributions for MFI scores.

MF was compared with CLR and WL_2Boost for unmatched data sets. However, for the gene expression data set, only WL_2Boost was used for comparison because

this data set has a large number of variables ($> 22,000$) and CLR has convergence problem for high dimensional data sets (Asafu-Adjei *et al.* (2017)).

Indian liver patient data set: This data set contains 416 instances labeled as liver patients (cases) and 145 instances labeled as non-liver patients (controls). We used age (discretized by 5 year intervals) and gender as matching variables and studied the effect of the remaining 8 exposure variables on liver disease: Total Bilirubin (x_1), Direct Bilirubin (x_2), Alkphos (x_3), SGPT (x_4), SGOT (x_5), Total Protiens (x_6), Albumin (x_7) and A/G ratio (x_8). Variable x_8 had 4 missing values which were replaced by the average of non-missing values. Using exact matching to match case and control instances by Age and Gender resulted in a data set with 153 matched case-control pairs.

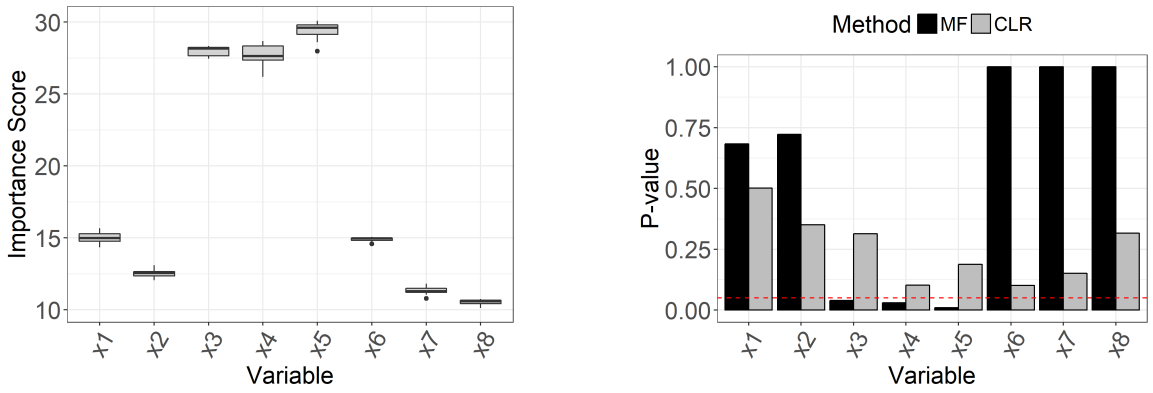
Figure 3.6a shows MFI scores for exposure variables and Figure 3.6b shows p-values of each exposure variable from MF and CLR. Variables x_3, x_4, and x_5 were selected by MF at significance level 0.05 and they all have received large MFI scores compared to other variables. However, CLR did not select any of these variables at this significance level. This difference can be due to existence of interactions between variables that CLR did not detect. To test if variable x_5 interacts with other exposure variables, a similar approach to Balasubramanian *et al.* (2014) was used. We fit the following two conditional logistic regression models

$$logit(p) = \beta_1 x_5$$

$$logit(p) = \beta_1 x_5 + \beta_2 z + \beta_3 x_5 z$$

where z is any other variable in the data set. A likelihood ratio test was performed at level 0.05 to compare the two nested models. Variable x_5 has significant interactions with variables x_1, x_2 and x_3. Also, the result from WL_2Boost differs from MF. WL_2Boost selected variable x_7 as an important variable, but this received a large

p-value by MF. This difference is possibly due to interactions between the variables which WL_2Boost has difficulty to detect.



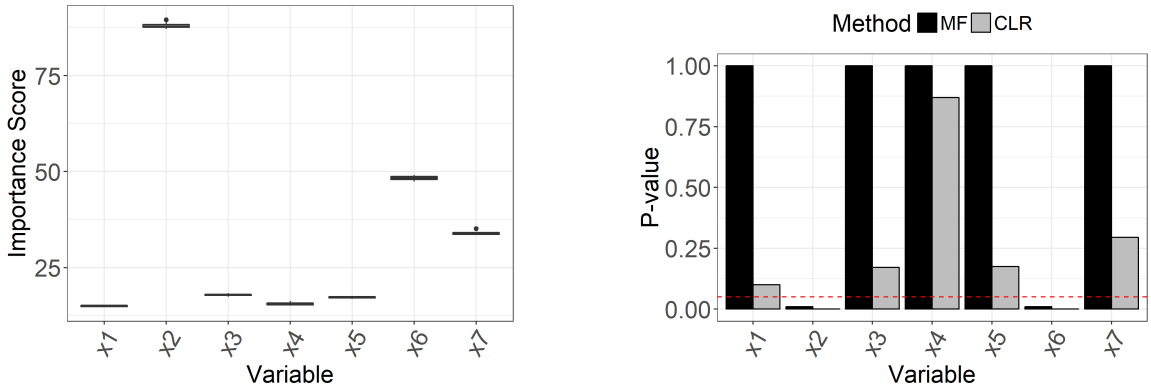
(a) MFI scores

(b) P-values from MF and CLR

Figure 3.6: Indian Liver Patient Data Set: (a) MFI Scores for Exposure Variables and (b) P-values from MF and CLR.

Pima indians diabetes data set: This data set contains diabetes diagnostic information for 268 women diagnosed with diabetes and 500 without. We matched instances based on age (discretized by 5 year intervals) and generated 241 pairs. The 7 exposure variables were number of times pregnant (x_1), Plasma glucose concentration (x_2), Diastolic blood pressure (x_3), Triceps skin fold thickness (x_4), 2-Hour serum insulin (x_5), Body mass index (x_6), and Diabetes pedigree function (x_7). Figure 3.7a shows MFI scores for exposure variables and Figure 3.7b compares p-values from MF and CLR. Here, the results obtained by MF, CLR, and WL_2Boost are similar and they all selected variables x_2 and x_6 as important variables. Also, we can observe in Figure 3.7a that both variables have relatively large MFI scores compared to other variables.

Statlog heart disease data set: The Statlog heart disease data set contains 120 and 150 instances, with and without heart disease, respectively. Exposure variables



(a) MFI scores

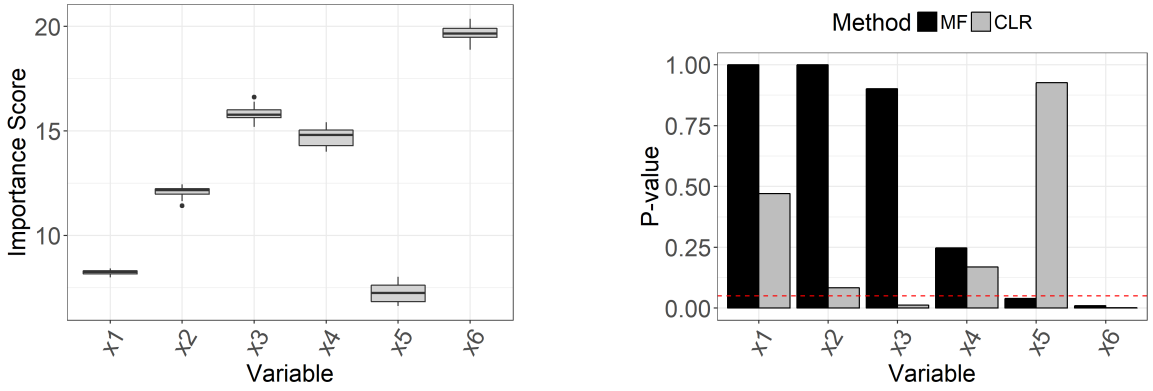
(b) P-values from MF and CLR

Figure 3.7: Pima Indians Diabetes Data Set: (a) MFI Scores for Exposure Variables and (b) P-values from MF and CLR.

of interest are Resting Blood Pressure (x_1), Serum Cholestoral (x_2), max heart rate (x_3), Oldpeak (x_4), Slope of peak ST segment (x_5), and Major vessels colored (x_6). Age (discretized by 5 year intervals) and gender were used to match controls to similar cases using the exact matching method. This resulted in 80 matched pairs.

Figure 3.8a shows MFI scores for exposure variables and Figure 3.8b shows p-values from MF and CLR. Variable x_6 was selected by both methods at significance level 0.05, but there exists several differences between the two methods. Variable x_5 was selected by MF, while CLR did not detect the effect of this variable. The reason why variable x_5 received a relatively small MFI score is that this variable has only 3 unique values and RF tends to get higher Gini importance score for numerical variables with several unique values (Strobl and Zeileis (2008)). From MFI scores in Figure 3.8a, we see that both x_3 and x_4 have relatively large and almost similar MFI scores, so they potentially have effects on the outcome. CLR only detected the effect of x_3, but MF did not detect the effect of x_3 nor x_4. If we look at p-values of x_r^{*0}, x_r^{*1}, and d_r^* for $r \in \{3, 4\}$, we observe that both d_3^* and d_4^* are significant at $\alpha = 0.05$,

thus, further work is required to refine the method and improve the power of MF. We also applied WL_2Boost on this data set and observed that only variable x_6 was selected.

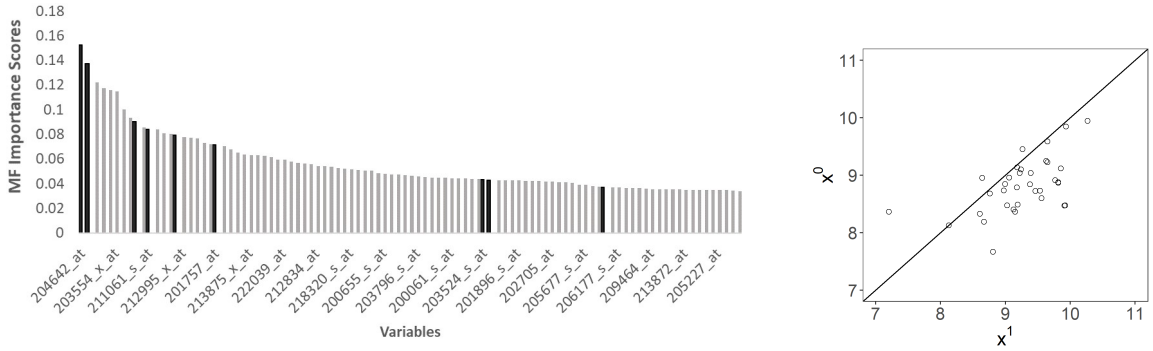


(a) MFI scores

(b) P-values from MF and CLR

Figure 3.8: Statlog Heart Disease Data Set: (a) MFI Scores for Exposure Variables and (b) P-values from MF and CLR.

Childhood Acute Lymphoblastic Leukemia Study: The childhood acute



(a)

(b)

Figure 3.9: (a) Average MFI Scores for the 100 Highest MFI Importance Variables. Genes Selected by WL_2Boost Are Shown in Dark Shade. (b) Scatter Plot of Control Versus Case for Gene 213166_x_at Indicates An Effect.

lymphoblastic leukemia study is a matched paired study design conducted by Bhojwani *et al.* (2006) to study underlying mechanisms leading to relapse. They analyzed gene expression profiles of 35 children who were diagnosed with childhood acute lymphoblastic leukemia and relapsed after therapy. This study used 35 pairs, each of which consisted of gene expression profiles in bone marrow of diagnosis and relapsed samples taken from the same patient. The study evaluated 22,283 gene expressions to identify genes which were differentially expressed between selected pairs. Although matching variable is not measured in this data set, we can still use MF to select important genes.

MF was conducted in 10 replicates and $50,000$ trees and compared with WL_2Boost. Variables (Gene expression profiles) were ranked from most (rank 1) to least (rank 22283) important based on MFI scores and important variables were selected by randomly assigned labels. At $\alpha = 0.01$, MF selected 284 variables which include 11 out of 12 variables selected by WL_2Boost and at $\alpha = 0.02$, all 12 variables were selected by MF. Figure 3.9a shows the average MFI scores over the 10 replicates of MF for the 100 highest MFI importance variables. Not all bars are labeled in the figure. Here 9 of the 12 variables which were selected by WL_2Boost are shown in dark shade (while all 12 are within the top 200 of MFI scores). WL_2Boost selected the top two MFI highest ranked variables. However, variables of rank 3 through 8 based on MFI scores, which also received small p-values, were not selected by WL_2Boost. Figure 3.9b is a scatter plot of the MFI rank 3 gene 213166_x_at and it illustrates that for many pairs $x^1 > x^0$ so that we see a strong effect based on our potential outcome approach. Also, Adewale *et al.* (2010) considered a classifier to distinguish the case and control samples. Based on a MF model from only the 10 MFI highest-ranked genes, we obtain a cross validated error rate of approximately 3%, substantially lower than the 23% error rate of the 12-variable model from WL_2Boost.

Also, we viewed a variable selected by WL_2Boost, but ranked lower by MF, such as 218561_s_at (155th). It can be shown that this variable has less predictive scatter plot (as measured by the difference between the number of pairs with $x^1 > x^0$ and $x^1 < x^0$) when viewed individually than approximately 900 other variables, so the lower rank is reasonable. Its rank of 155 is higher than the scatter plot measure would indicate, presumably from a role of this variable in interactions captured by MFI scores. Still, our model ranked many other genes higher and this gene was not needed for a good prediction.

3.5 Conclusions

We presented Matched Forest (MF), a machine learning algorithm for variable selection in high dimensional matched case-control data set. The method differs substantially from previous approaches and is developed to detect complex effects. Data is transformed so that advantages of supervised learners can be used to identify important variables. MF also uses a modified variable importance measure for variable selection. MF is both conceptually simple and easy to apply with widely available software tools.

We compared the performance of MF to alternative approaches including CLR, RF and WL_2Boost using simulated and biomedical data sets. The simulation studies demonstrate the effectiveness of MF to detect important variables with interaction effects. Also, the analysis on biomedical data sets shows that results from MF can be different from alternative approaches because of its ability to detect complex interaction effects.

3.6 Supporting Information: Simulation Studies

The full description of simulations and their results are presented here. Data sets are generated with different conditions defined by effect type, the size of the effect, number of exposure variables and number of strata to test the performance of Matched Forest (MF) in detecting important variables. MF was compared with Random Forest(RF) and Conditional Logistic Regression (CLR) based on ROC curve which plots true positive rate (TPR) versus false positive rate (FPR) at different values of significance level α. For the simulations in section 3.6.1 where there is no effect, the three methods were compared based on their FPR at different values of α. We also show the MFI scores for the top 20 most important exposure variables and for all 5 matching variables in our simulation studies.

3.6.1 Null Scenario

For the null scenario, we simulated data sets with no effect of exposure variables. Two simulation experiments were designed to generate data sets with no effects. In the first experiment, all variables (matching and exposure) were simulated independently and in the second experiment, a matching variable (v_1) is simulated to be associated with the exposure variable (x_1). We compared the performance of MF, RF and CLR when $N = 600$ and $R = 100$ for both simulation experiments. For the first experiment with independent variables, we set $\mu_r = 0$ for $r = 1, 2, \ldots, R$. Figure 3.10 shows the results for the simulation experiment with independent variables. Figure 3.10a compares FPR values from MF, RF, and CLR at different values of α for this experiment. We can see that MF and CLR both have small and similar FPR and $\alpha < 0.3$ guarantees that MF has smaller FPR than RF. Also, MFI scores of matching and exposure variables do not show any variable with significantly higher

score than the other matching and exposure variables. (Figures 3.10b, 3.10c). We also considered the effect of N and R in the performance of MF and observed that FPR of MF is similar for different values of N and R that we tested and almost equal to the significance level α (Figure 3.12).

For the second simulation experiment, x_1^0 and x_1^1 are generated from matching variable v_1 according to v_1+t where t follows $N(0,1)$. All other variables are simulated in the same fashion as the first experiment with $\mu_r = 0$. Figure 3.10 shows the results for the simulation experiment with matching associated with an exposure. Figure 3.11a compares FPR values from MF, RF, and CLR at different values of α for this experiment. We can see that MF and CLR both have small and similar FPR and $\alpha < 0.3$ guarantees that MF has smaller FPR than RF. Also, MFI scores of matching and exposure variables do not show any variable with significantly higher score than the other matching and exposure variables. (Figures 3.11b and 3.11c). We also considered the effect of N and R in the performance of MF and observed that FPR of MF is similar for different values of N and R that we tested and almost equal to the significance level α. (Figure 3.13) Therefor, MF behaves comparably to CLR in terms of FPR when all variables are simulated with no effect. Also, MF and CLR perform better than RF only when $\alpha < 0.3$.

3.6.2 Linear Exposure Effect

Here, exposure variable x_1 is simulated to have a negative effect. In particular, we set $\mu_1 = -1$ for $N/2$ instances and to generate some noise, $\mu_1 = 0$ for the remaining instances, and $\mu_r = 0$ for $r = \{2, \ldots, R\}$. Figure 3.14a shows ROC curves from MF, RF, and CLR, and Figures 3.14b and 3.14c show MFI scores for matching and exposure variables respectively. From Figure 3.14a, we can see that both MF and CLR are accurate in terms of selecting important variable x_1 and their TPR is always 1

at any given FPR. However, RF has smaller TPR than MF and CLR at any given FPR. We can also observe that variable x_1 has substantially higher MFI score than the other exposure variables (Figure 3.14c) and all 5 matching variables have small and almost identical MFI scores (Figure 3.14b).

We also tested the effect of number of pairs (N), number of exposure variables (R), and μ_1 for instances with negative effect ($\mu_1(-)$) in the performance of MF and observed that MF is accurate in identifying variable x_1 as important in ranges of values that we tested and its TPR is equal to 1 at any given FPR (Figure 3.15). Therefore, both MF and CLR perform comparably and better than RF in identifying variables with linear effect.

3.6.3 Non-Linear Exposure Effect

Here, x_1 has an effect that changes with the value of x_1^0, that is, x_1 has a positive effect when $x_1^0 < 25$ and has a negative effect when $x_1^0 > 25$. We use a uniform distribution to generate x_1^0 such that the first $N/6$ instances are between 1 and 25, the next $N/3$ instances are between 25 and 50 and the remaining instances are between 1 and 50. To generate x_1^1, $\mu_1 = 2$ for the first $N/6$ instances (positive effect), $\mu_1 = -1$ for the next $N/3$ instances (negative effect), $\mu_1 = 0$ for the remaining $N/2$ instances. Also, $\mu_r = 0$ for $r = \{2, \ldots, R\}$. Figure 3.16a shows ROC curves from MF, RF, and CLR and Figures 3.16b and 3.16c show MFI scores for matching and exposure variables respectively for data sets with a non-linear effect, $N = 600$ and $R = 100$. Figure 3.16a shows that MF dominates the other two methods as it always has higher TPR at any given FPR. Also, variable x_1 has received a substantially higher MFI score than the other exposure variables (Figure 3.16c) and MFI scores of matching variables are all small and almost equal and no matching variable with unusually high score is seen in Figure 3.16b as expected.

We also evaluated the performance of MF for different values of number of pairs (N), number of exposure variables (R), and μ_1 for instances with positive and negative effects ($\mu_1(+), \mu_1(-)$). Figure 3.17 shows the results from this sensitivity analysis for $N \in \{300, 600, 900\}$, $R \in \{20, 100, 150\}$, $(\mu_1(+), \mu_1(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$ respectively. The performance of MF improves by increasing the number of pairs from 300 to 900. MF also is more accurate when number of variables is small and its performance slightly drops as the number of variables increases. Also, as the effect size ($|\mu_1|$) increases, MF will be more accurate in identifying the correct effect. Therefore, MF performs better than both CLR and RF in identifying the effect of variables with non-linear effects, and the performance of MF in variable selection improves with larger number of strata, smaller number of exposure variables, and an increase in effect size of the important variable.

3.6.4 Matching-Exposure Interaction

Here, data sets are generated with an interaction between exposure variable x_1 and matching variable v_1 with no other effects in a similar fashion as described in section 4.1 in the manuscript. We compared the performance of MF with CLR and RF in Figure 8 of the manuscript. In this section, we show how the performance of MF is affected by change in number of pairs (N), number of exposure variables (R), and μ_1 for instances with positive and negative effects ($\mu_1(+), \mu_1(-)$). Figure 3.18 shows the results from this sensitivity analysis for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_1(+), \mu_1(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF is accurate in identifying the interaction effect between v_1 and x_1 and its TPR is near or equal to 1 at all tested values for N and R. Also, the performance of MF improves as the effect size ($|\mu_1|$) increases. Therefore, MF is effective in identifying an interaction between a matching and an exposure variable, and its performance improves as effect size increases.

3.6.5 *Exposure-Exposure Interaction*

An interaction between exposure variables x_1 and x_2 is generated, without individual effects. That is, $\mu_1 = \mu_2 = 2$ for the $N/4$ instances, $\mu_1 = \mu_2 = -1$ for $N/2$ instances, and to create some noise, $\mu_1 = \mu_2 = 0$ for the remaining $N/4$ instances. Other exposure variables are generated with $\mu_r = 0$ for $r = \{3, \ldots, R\}$. Figure 3.19a shows ROC curves from MF, RF, and CLR and Figures 3.19b and 3.19c show MFI scores for matching and exposure variables respectively for data sets with exposure-exposure interaction, $N = 800$ and $R = 100$. From Figure 3.19a, we can see that MF outperforms RF and CLR in identifying the correct effects because its TPR is almost always higher than TPR of RF and CLR at any given value of FPR. Also, MFI scores in Figure 3.19b and Figure 3.19c show that both x_1 and x_2 have significantly higher scores than other exposure variables and there is no matching variable whose MFI score is significantly higher than other matching variables.

We also considered the effect of number of pairs (N), number of exposure variables (R), and μ_2 for instances with positive and negative effects $(\mu_2(+), \mu_2(-))$ on the performance of MF. Figure 3.20 shows the results from this sensitivity analysis for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_2(+), \mu_2(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF has TPR equal or very close to 1 at any given FPR for all values of N and R that we tested so its performance is not sensitive to changes in these two parameters. Also, as the effect size ($|\mu_2|$) increases, MF performs better in identifying variables in an exposure-exposure interaction. Therefore, MF performs better than both CLR and RF in identifying interaction effects between two exposure variables, and the performance of MF in variable selection improves with an increase in the strength of the interaction effect.

3.6.6 Exposure-Exposure-Exposure Interaction

Data sets with a three-way interaction effect between exposure variables x_1, x_2, and x_3 and with no effects for other variables are generated for this simulation study. Table 3.1 describes how the interaction is designed. We indicate pairs with negative and positive effect of each case variable with $-$ and $+$, respectively. For example, the first row in Table 3.1 indicates that $N/8$ instances are generated with positive effect for x_1^1, negative effect for x_2^1, and positive effect for x_3^1. Number of pairs with negative and positive effects and μ_r are selected such that no effect is observed from one of the variables or a combination of two, but all three variables together show an effect. In our simulations $\mu_r(+) = 2|\mu_r(-)|$ for $r \in \{1, 2, 3\}$. We also added some noise to data sets by generating $N/8$ pairs with $\mu_1 = \mu_2 = \mu_3 = 0$.

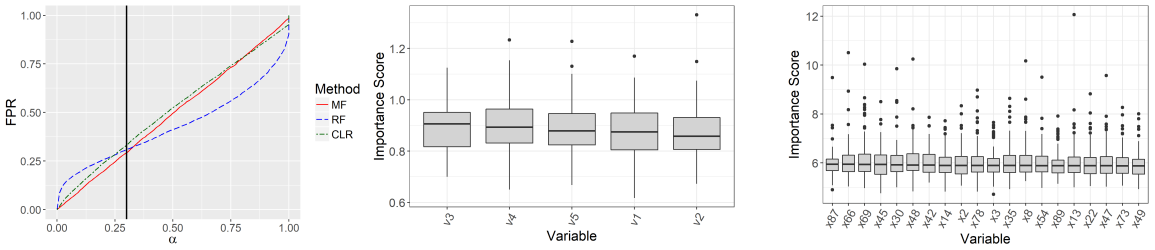
Table 3.1: Design Table for Exposure-exposure-exposure Interaction Simulation.

x_1^1 Case 1	x_2^1 Case 2	x_3^1 Case 3	Number of instances
+1	-1	+1	$N/8$
+1	+1	-1	$N/8$
-1	-1	-1	$3N/8$
-1	+1	+1	$N/8$

Figure 3.21a shows ROC curves from MF, RF, and CLR, Figure 3.21b shows MFI scores of matching variables and Figure 3.21c shows MFI scores of exposure variables for data sets with exposure-exposure-exposure interaction, $N = 800$ and $R = 100$. There can be seen in Figure 3.21a that MF has better performance in identifying important variables compared to RF and CLR because its ROC curve dominates the ROC curves of RF and CLR. Also, MFI scores for exposure variables in Figure 3.21c show that x_1, x_2 and x_3 have significantly higher scores than other exposure variables

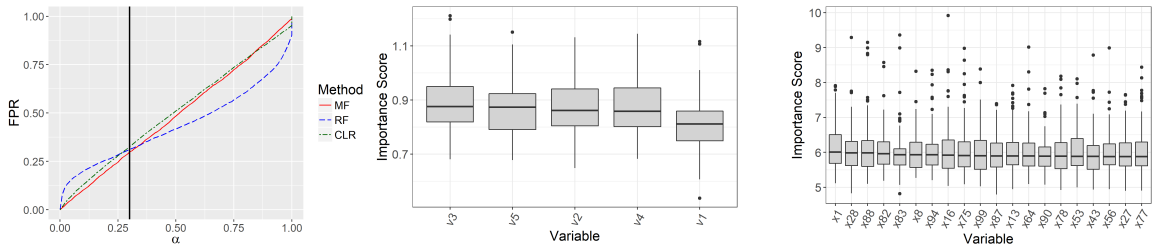
and MFI scores of matching variables in Figure 3.21b do not indicate any variable with significantly large MFI score than other variables.

We also tested the effect of number of pairs (N), number of exposure variables (R), and μ_3 for instances with positive and negative effects ($\mu_3(+), \mu_3(-)$) on the performance of MF. Figure 3.22 shows the results for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_3(+), \mu_3(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. The performance of MF is good for all values of N and R that we tested and its TPR is equal or very close to 1 at any given FPR. Also, as the effect size ($|\mu_3|$) increases, the performance of MF improves in selecting variables in an exposure-exposure-exposure interaction. Therefore, MF performs better than both CLR and RF in identifying interaction effects between three exposure variables, and the performance of MF in variable selection improves with an increase in the strength of the interaction effect.



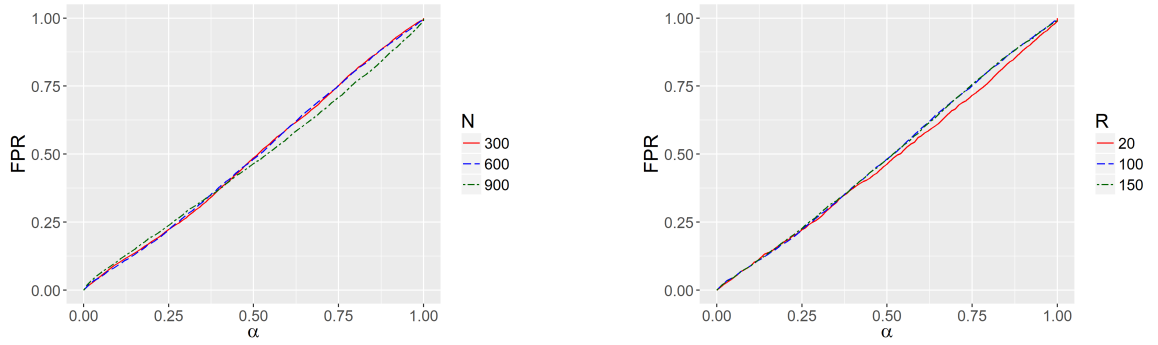
(a) FPR curve (b) MFI scores: matching variables (c) MFI scores: exposure variables

Figure 3.10: Null Scenario With Independent Variables: (a) Comparison Between the Performance of MF, RF, and CLR in Terms of FPR, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated with $N = 600$ and $R = 100$. MF and CLR Have Small and Similar FPR and If $\alpha < 0.3$, MF Has Smaller FPR than RF. MFI Plots Do Not Indicate Any Variable With Unusual High Score.



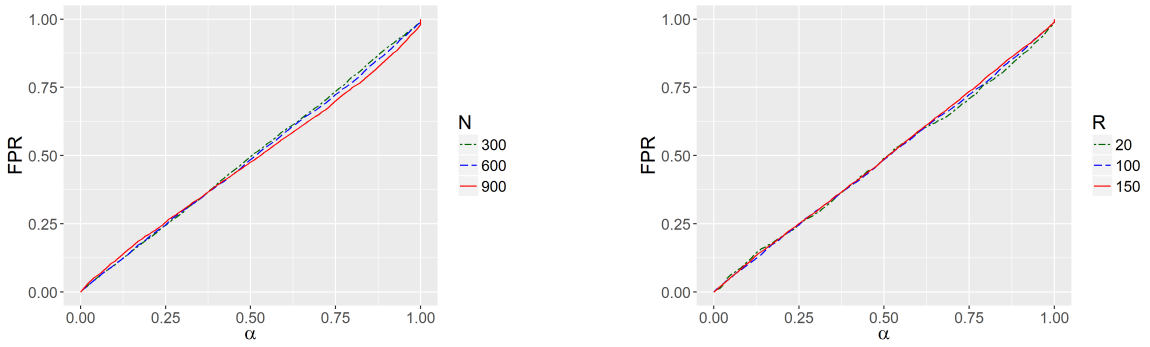
(a) FPR curve (b) MFI scores: matching variables (c) MFI scores: exposure variables

Figure 3.11: Null Scenario with Matching Associated with Exposure: (a) Comparison Between the Performance of MF, RF, and CLR in Terms of FPR, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated With $N = 600$ and $R = 100$. MF and CLR Have Small and Similar FPR and If $\alpha < 0.3$, MF Has Smaller FPR Than RF. MFI Plots Do Not Indicate Any Variable With Unusual High Score.



(a) Effect of N (b) MFI scores

Figure 3.12: Null Scenario with Independent Variables: Effect of Number of Pairs (N) and Number of Exposure Variables (R) on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$ and $R \in \{20, 100, 150\}$. FPR of MF Is Small at Different Values of α and Is Not Sensitive to the Range of Values That We Tested for N and R.



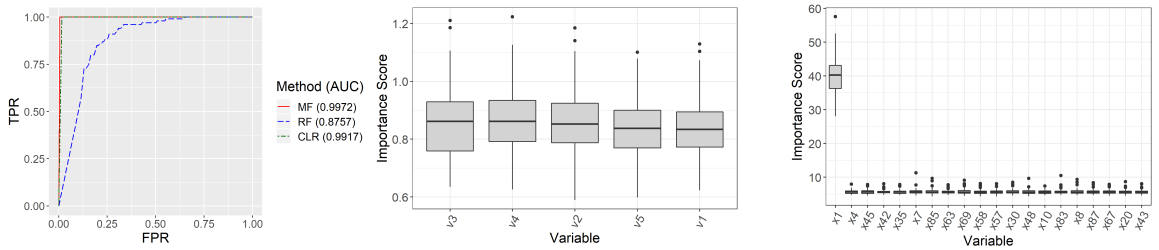
(a) Effect of N

(b) MFI scores

Figure 3.13: Null Scenario With Matching Associated With Exposure: Effect of Number of Pairs (N) and Number of Exposure Variables (R) on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$ and $R \in \{20, 100, 150\}$. FPR of MF Is Small at Different Values of α and Is Not Sensitive to the Range of Values That We Tested for N and R.

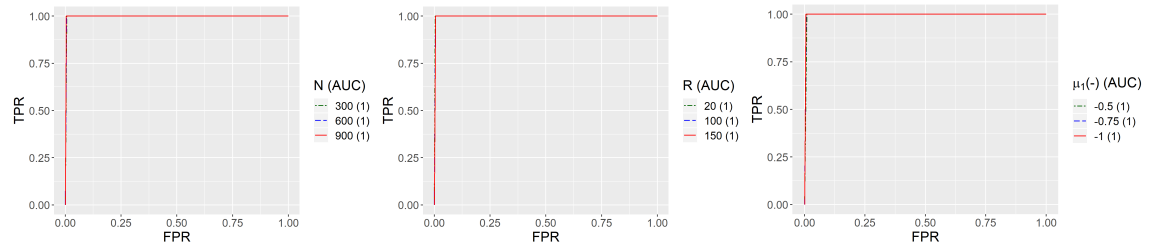
3.6.7 Summary of simulation results

We observed in our simulations that MF performs better than CLR in identifying interaction effects between exposure variables and matching and exposure variables, however, their performance is comparable when the important variable has a linear effect. Also, MF performs better than RF for different effect types including linear, non-linear and interactions. The variable selection performance of MF generally improves with larger number of strata, larger effect size and smaller number of exposure variables.



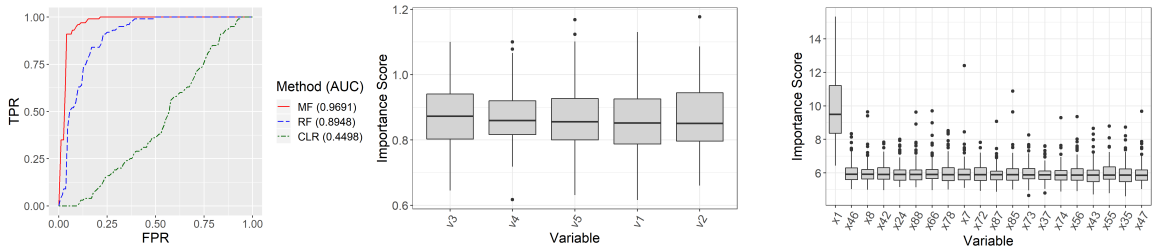
(a) ROC curves (b) MFI scores: matching variables (c) MFI scores: exposure variables

Figure 3.14: Linear Exposure: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated with a Linear Effect of x_1, $N = 600$ and $R = 100$. MF and CLR Are Accurate in Detecting the Linear Effect of x_1 and Their Roc Curve Dominates RF. MFI Plots Show That x_1 Has Substantially Higher MFI Score than the Other Exposure Variables and Matching Variables Have Small and Almost Identical MFI Scores.



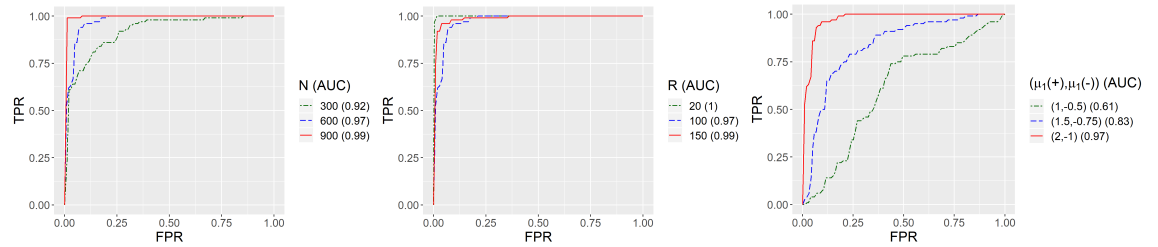
(a) Effect of N (b) Effect of R (c) Effect of μ_1

Figure 3.15: Linear Exposure: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_1 for Instances With Negative Effect ($\mu_1(-)$) on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$, $R \in \{20, 100, 150\}$, and $\mu_1(-) \in \{-0.5, -0.75, -1\}$ Respectively. MF Selects Variable x_1 As Important in Different Ranges of Values That We Tested and Its TPR Is Always Equal to 1 at Any Given Value of FPR.



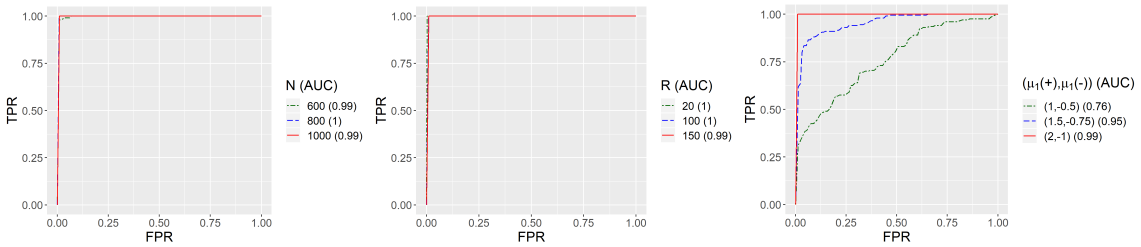
(a) ROC curves (b) MFI scores: matching variables (c) MFI scores: exposure variables

Figure 3.16: Non-linear Exposure: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables. Results Are Shown for Data Sets Simulated with Non-linear Exposure Effect of x_1, $N = 600$ and $R = 100$. MF Performs Better than RF and CLR in Identifying the Non-linear Effect of x_1. MFI Plot Shows That x_1 Has Substantially Higher MFI Score than the Other Exposure Variables and All Matching Variables Have Small and Similar MFI Scores.



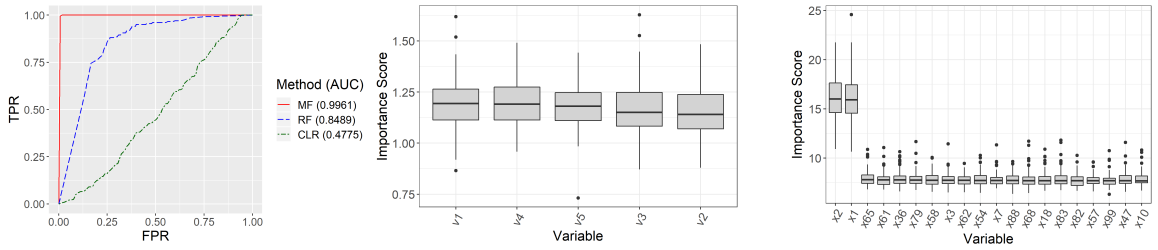
(a) Effect of N (b) Effect of R (c) Effect of μ_1

Figure 3.17: Non-linear Exposure: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_1 for Instances with Positive and Negative Effects $(\mu_1(+), \mu_1(-))$ on MF Performance. Results Are Shown for $N \in \{300, 600, 900\}$, $R \in \{20, 100, 150\}$, $(\mu_1(+), \mu_1(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$ Respectively. The Performance of MF Improves with Larger Number of Pairs, Effect Size and Smaller Number of Variables.



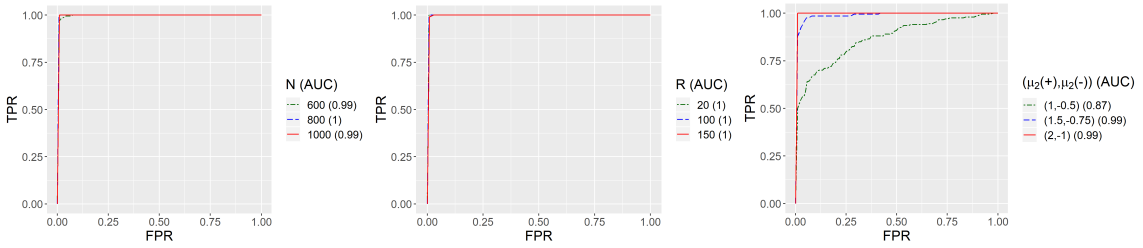
(a) Effect of N (b) Effect of R (c) Effect of μ_1

Figure 3.18: Matching-exposure Interaction: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_1 for Instances with Positive and Negative Effects ($\mu_1(+), \mu_1(-)$) on MF Performance in Identifying a Matching-exposure Interaction. Results Are Shown for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, and $(\mu_1(+), \mu_1(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF Selects Important Variables Accurately with TPR near or Equal to 1 at Any given Value of FPR for All Ranges of Values That We Tested for N and R and Its Performance Improves with Larger Effect Size.



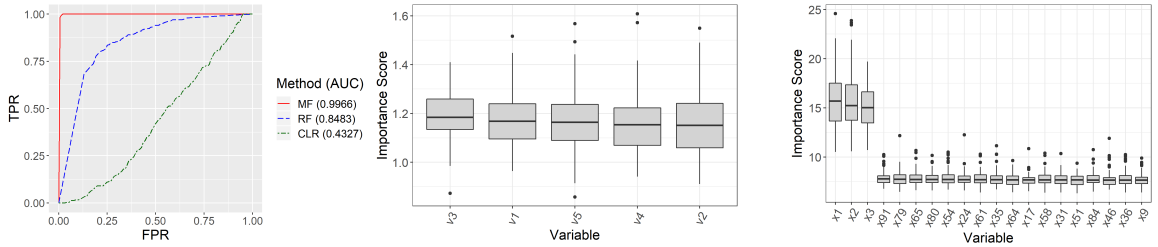
(a) ROC curve (b) MFI scores: matching variables (c) MFI scores: exposure variables

Figure 3.19: Exposure-exposure Interaction: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables for Data Sets Simulated with Exposure-exposure Interaction Between x_1 and x_2, $n = 800$ and $r = 100$. MF Performs Better than RF and CLR in Identifying the Correct Effect. MFI Plots Show That Both x_1 and x_2 Have Considerably Higher MFI Scores than the Other Exposure Variables and There Is No Matching Variable with Significantly Higher MFI Score than Others.



(a) Effect of N (b) Effect of R (c) Effect of μ_2

Figure 3.20: Exposure-exposure Interaction: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_2 for Instances with Positive and Negative Effects ($\mu_2(+), \mu_2(-)$) on MF Performance in Identifying An Exposure-exposure Interaction. Results Are Shown for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_2(+), \mu_2(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF Is Almost Accurate in All the Settings That We Tested for N and R and Its Performance Improves for Larger Effect Size.

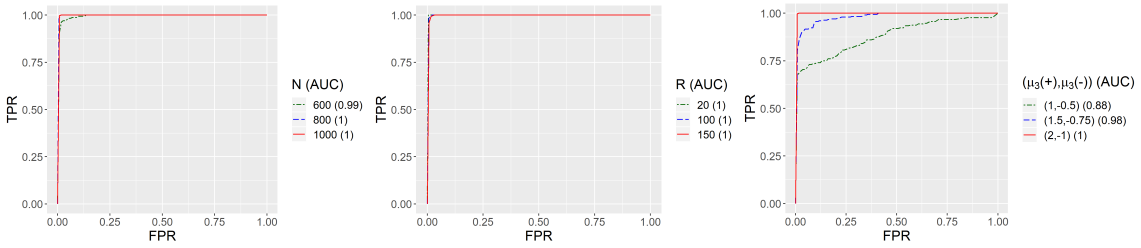


(a) ROC curve

(b) MFI scores: matching variables

(c) MFI scores: exposure variables

Figure 3.21: Exposure-exposure-exposure Interaction: (a) Comparison Between the Performance of MF, RF, and CLR in Variable Selection Accuracy, (b) MFI Scores of Matching Variables and (c) MFI Scores of Exposure Variables for Data Sets Simulated With Exposure-exposure-exposure Interaction Between x_1, x_2 and x_3, $N = 800$ and $R = 100$. The Performance of MF Is Better Than RF and CLR in Identifying the Correct Effects. MFI Plots Show That Variables x_1, x_2 and x_3 Have Received Considerably Higher MFI Scores Than the Other Exposure Variables and MFI Scores for Matching Variables Do Not Indicate Any Variable With Significantly Larger Score Than Others. All the Settings That We Tested for N and R and Its Performance Improves for Larger Effect Size.



(a) Effect of N (b) Effect of R (c) Effect of μ_3

Figure 3.22: Exposure-exposure-exposure Interaction: Effect of Number of Pairs (N), Number of Exposure Variables (R), and μ_3 for Instances with Positive and Negative Effects $(\mu_3(+), \mu_3(-))$ on MF Performance in Identifying An Exposure-exposure-exposure Interaction Effect. Results Are Shown for $N \in \{600, 800, 1000\}$, $R \in \{20, 100, 150\}$, $(\mu_3(+), \mu_3(-)) \in \{(1, -0.5), (1.5, -0.75), (2, -1)\}$. MF Performs Well for All Values of N and R That We Tested and Its Performance Improves as Effect Size Increases.

Chapter 4

ENHANCEMENTS OF MATCHED FOREST

4.1 Introduction

Matched case-control analysis is commonly used in a wide range of applications to remove the effect of confounding variables in identifying important variables and building predictive models. Case and control correspond to units with and without the condition of interest respectively. Case and control instances are grouped into a stratum based on some matching variables and a number of exposure variables are studied for their effect on the condition of interest. In clinical applications, matching is usually done on demographic variables such as age and gender to remove the effect of confounding variables which can lead to spurious results. For example, Balasubramanian *et al.* (2014) matched subjects with and without cardio-vascular disease (case and control) based on age, gender, race and severity of coronary artery disease to identify biomarkers associated with the disease.

There are several variable selection and prediction approaches for analysing matched case-control data sets. Conditional logistic regression (CLR) (Hosmer and Lemeshow (2000)) and its variants for high-dimensional data sets (Balasubramanian *et al.* (2014), Asafu-Adjei *et al.* (2017), and Qian *et al.* (2014)) have been commonly used for variable selection from matched case-control data sets. Another commonly used method for matched case-control data set is generalized linear mixed models (GLMM) for binary outcomes. For example, Szyszkowicz (2006) used GLMM to evaluate the association between air pollution and health outcomes. Two modified boosting algorithms were also proposed by Adewale *et al.* (2010) for data sets with

correlated binary outcome and recently Stanfill *et al.* (2019) presented a conditional classification method to account for matched structure of data in distinguishing between case and control units. However, all of these algorithms struggle to detect non-linear and interaction effects, particularly when data set is high-dimensional.

Recently, Shomal Zadeh *et al.* (2020) proposed Matched Forest, a supervised machine learning method for variable selection in high-dimensional matched case-control data sets. This method converts a matched case-control data set to a supervised setting which controls for its matched structure and then applies Random Forest (RF) on the transformed data set to compute variable importance score and select important variables. Shomal Zadeh *et al.* (2020) evaluated the variable selection accuracy of MF in high-dimensional data sets with different types of effects. They observed that MF has a significantly better variable selection performance than competing algorithms, including CLR, Boosting Weighted L_2 Loss (WL_2Boost) (Adewale *et al.* (2010)), and unmatched RF.

Here, we propose three enhancements of MF to improve its performance in high-dimensional setting, evaluate its classification accuracy and estimate the effect of important variables. First, a regularized version of MF is proposed to improve its variable selection accuracy in extremely high-dimensional matched case-control data sets. Our regularized version of MF which we refer to as Weighted Matched Forest (WMF) is motivated by Basu *et al.* (2018). It sequentially grows a feature-weighted version of MF to reduce the dimensionality and focus on highly informative variables. Second, we propose a prediction method based on MF to classify matched pairs into case-control or control-case. For the classification approach, we use the predicted labels by MF for observed and counterfactual pairs to predict a label for an unlabeled pair of data. Third, we propose a method to estimate the effect of important exposure variables selected by MF.

In Section 4.2, we present background on classification algorithms for matched case-control data sets and Matched Forest algorithm. Section 4.3 proposes methods for three enhancements of MF, including Weighted Matched Forest (WMF), classification of matched pairs and effect estimation. Section 4.4 presents the results and section 4.5 provides the conclusion.

4.2 Background

4.2.1 Classification of Matched Case-Control Data Sets

Majority of classification algorithms for matched case-control data sets are based on Conditional Logistic Regression (CLR). CLR and its variants for high-dimensional setting (Balasubramanian *et al.* (2014), Asafu-Adjei *et al.* (2017), and Qian *et al.* (2014)) predict the conditional probability that the first member in a stratum is a case given the assumptions: (1) there is only one case in each stratum and (2) logistic regression is the correct model to predict outcomes of subjects in the stratum. Formally, let $(x^1(i), x^0(i))$ denote feature vectors and $(y^1(i), y^0(i))$ be case-control status corresponding to case and control subjects in pair i respectively, such that $y^j(i)$ takes 1 for cases and 0 for controls for $j \in \{0, 1\}$. CLR computes the conditional probability as follows:

$$p(y^1(i) = 1 | y^1(i) + y^0(i) = 1, x^1(i), x^0(i)) \tag{4.1}$$

Thus, for matched pairs, CLR predicts the probability that a pair of subjects is either case-control or control-case, because only one instance in a pair can be a case. CLR and its variants for high-dimensional setting use a linear model which requires interaction terms (products of two or more variables) to assess their effect. Thus, they are not suitable for high-dimensional matched case-control data sets with complex relationships among variables.

Matched case-control data sets have been also analyzed by methods which are not based on CLR. Adewale *et al.* (2010) proposed two variants of boosting algorithm for the classification of correlated binay data sets and selecting important variables. The first method is Boosting Weighted L_2 Loss (WL_2Boost) which uses the gradient boosting algorithm with a modified loss function that handles the correlation among subjects within a stratum. The second method, Penalized Quasi-Likehood Boosting algorithm (PQLBoost), modifies the likelihood function in the boosting algorithm to make it suitable for matched case-control data sets. Both methods classify each instance as either case or control, without any assumption on the number of cases within each stratum. However, they are both linear methods and are not able to handle non-linear and interactions effects.

Recently, Stanfill *et al.* (2019) proposed a data transformation approach to generalize classification algorithms to matched case-control data sets. Specifically, they center each strata by the mean values of exposure and map the exposure value of each unit to its difference from the center. This method does not handle dependency among units within a stratum, which is recommended by statistical principles. It breaks each stratum into multiple instances which are known to be dependent. However, our proposed method construct data sets where instances are independent. We show in our experiments in Chapter 5 that variable selection performance of our method is better than the method proposed by Stanfill *et al.* (2019), especially when informative variables have nonlinear or interaction effects.

4.2.2 Matched Forest (MF)

MF (Shomal Zadeh *et al.* (2020)) is applied on matched pairs to identify exposure and matching variables which are important in distinguishing between case and control subjects. MF consists of two steps, including a transformation of matched

pairs to a supervised setting based on potential outcome framework of causal inference and a classifier which inherently detects interactions involving both matching and exposure variables in high-dimensional setting. A wide range of classifiers can be used in the second step. Shomal Zadeh *et al.* (2020) chose RF due to its ability to handle high-dimensionality, mixed variables (numerical and categorical), interactions and non-linear effects.

In the data transformation step, MF creates new case and control variables x_r^{*1} and x_r^{*0} for each exposure variable x_r with $2N$ instances as

$$x_r^{*k}(i) = \begin{cases} x_r^k(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ x_r^{1-k}(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{4.2}$$

for $r \in \{1, 2, \ldots, R\}$ and $k \in \{0, 1\}$. That is, for the first N rows, the case and control values match the original pairs, but for the second N rows (referred as counterfactual), the values of case and control are interchanged within each pair. Variables d_r^* are also created for each numerical exposure variable as

$$d_r^* = x_r^{*1} - x_r^{*0} \tag{4.3}$$

to help identify the correct effect. Variables $v_1^+, v_2^+, \ldots, v_M^+$ are also created by extending the original matching variables as

$$v_m^+(i) = \begin{cases} v_m(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ v_m(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{4.4}$$

to help identify the interaction effects between matching and exposure variables. A label is also defined for each pair as

$$y(i) = \begin{cases} 0 & \text{for } i \in \{1, 2, \ldots, N\}, \\ 1 & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{4.5}$$

to distinguish between observed values and counterfactuals. If all exposure variables are numerical, this transformation will create new matched case-control data set D^* with $2N$ instances and $M + 3R + 1$ columns.

In the second step of MF, RF is applied on D^* to distinguish between observed pairs and their counterfactuals and identify important matching and exposure variables based on a new variable importance score denoted as MFI. The new MFI score is computed from the variable importance score of RF based on mean decrease in Gini information gain denoted as VI. The MFI score of exposure variable x_r is computed as

$$MFI(x_r) = VI(x_r^{*0}) + VI(x_r^{*1}) + VI(d_r^*) \tag{4.6}$$

for $r \in \{1, 2, \ldots, R\}$ and the MFI score of matching variable v_m is computed as

$$MFI(v_m^+) = VI(v_m^+) \tag{4.7}$$

for $m \in \{1, 2, \ldots, M\}$

4.3 Weighted Matched Forest, Classification And Effect Estimation

Section 4.3.1 proposes WMF which is an enhancement of MF to increase its power in identifying important variables. Section 4.3.2 explains how MF can be used for classification of matched pairs. In section 4.3.3, we propose a method to estimate the effect of important variables selected by MF.

4.3.1 Weighted Matched Forest

Consider a matched case-control data set D with N strata consisting of one case and one control, R exposure variables denoted by $\{x_1, x_2, \ldots, x_R\}$ and M matching variables denoted by $\{v_1, v_2, \ldots, v_m\}$. Let D^* be the transformed data set obtained by

MF according to Section 4.2.2 with new variables x_r^{*0}, x_r^{*1} and d_r^* for $r \in \{1, 2, \ldots, R\}$ and v_m^+ for $m \in \{1, 2, \ldots, M\}$.

Weighted Matched Forest (WMF) adaptively regularizes MF to focus on highly important variables for splits. Let $w = \{w_1, w_2, \ldots, w_p\}$ be a vector of non-negative weights corresponding to p variables in the transformed data set (D^*), and RF(w) be the RF algorithm built on D^* using feature weights w. In RF(w), instead of random sampling of variables at each node, variable j is selected proportional to w_j so that variables with higher weights have more chance to be selected for splits. In the original Breiman's RF algorithm (Breiman (2001)), all variables have similar weights equal to $1/p$.

Weighting can be done by measuring the ability of each variable to predict the outcome. For example, Basu *et al.* (2018) used variable importance score of RF to assign weights to each variable in unmatched data sets. To make the weighting suitable for matched setting, we should account for the dependency between case, control, and difference variables associated with an exposure in D^*, because if an exposure variable is important, all of its associated case, control, and difference variables help detect its effect. WMF modifies the weighting method proposed by Basu *et al.* (2018) for matched data sets.

The WMF algorithm updates the weights in a number of iterations. Given an iteration number T, WMF builds RF(w) iteratively on the transformed data set D^* with feature weights equal to w. The first iteration of WMF is equivalent to MF; WMF starts with equal weights $w^1 = \{1/p, 1/p, \ldots, 1/p\}$ for all p variables in D^* and the variable importance scores of RF(w^1) are stored as VI^1 to update the weights for the next iteration. This is repeated until iteration T, that is, in each iteration t, we compute w^t from the variable importance scores at previous iteration VI^{t-1} and build RF(w^t) to compute the importance scores VI^t. The MFI scores for exposure

and matching variables are computed in the final iteration of WMF (T) according to Equations 4.6 and 4.7 respectively.

Equations 4.8 and 4.9 show how weights are determined for matching and exposure variables, respectively, in matched case-control data sets with one case and one control in each stratum. Basically, the weight of a matching variable v_m^+ at iteration t is equal to its VI score at previous iteration of RF and the weights of variables corresponding to an exposure variable x_r, including x_r^{*0}, x_r^{*1}, and d_r^* are equal to the average of their VI scores at previous iteration of RF.

$$w^t(v_m^+) = VI^{t-1}(v_m^+) \tag{4.8}$$

$$w^t(x_r^{*0}) = w^t(x_r^{*1}) = w^t(d_r^*) = \frac{VI^{t-1}(x_r^{*0}) + VI^{t-1}(x_r^{*1}) + VI^{t-1}(d_r^*)}{3} \tag{4.9}$$

In our simulations, we tested different weighting methods for exposure variables including summation and maximum of $VI^{t-1}(x_r^{*0})$, $VI^{t-1}(x_r^{*1})$ and $VI^{t-1}(d_r^*)$, but we observed that WMF which uses the average of these scores as weights in each iteration of WMF outperforms the other two weighting methods in selecting the important matching and exposure variables. The reason why we use an equal weight for x_r^{*0}, x_r^{*1}, and d_r^* in each iteration of WMF is that if variable x_r is important, all of the three corresponding variables can help classify observed and counterfactual pairs, and by giving all the same weight, we allow all three of them to have the same chance to be selected at a node of Random Forest. As it will be shown in our simulations in section 4.4.1, this approach improves the variable selection performance of MF in identifying important variables.

4.3.2 Matched Forest for Classification of Matched Pairs

In this section, we explain how MF is applied for the classification of matched pairs. Given an unlabeled matched pair i, the objective is to map each instance in

the pair to either case or control given the assumption that there is only one case in the pair. Thus, the problem is equivalent to predicting the label of a pair y_i as either case-control ($y(i) = 0$) or control-case ($y(i) = 1$). Let $p(y(i) = 0)$ and $p(y(i) = 1)$ denote the predicted probabilities that the label of pair i is 0 and 1 respectively. Also, let $\Delta(i)$ denote the margin for pair i which measures how certain a classifier is in its prediction for pair i. The margin for pair i is defined as

$$\Delta(i) = |p(y(i) = 0) - p(y(i) = 1)| \tag{4.10}$$

Larger values of margin indicate that classifier is more certain about its prediction.

Our method obtains the label of pair i by comparing the predicted labels and margin for the observed pair i and its corresponding counterfactual pair denoted by $\tilde{i}$. The feature vectors corresponding to the counterfactual pair $\tilde{i}$ is obtained by permuting the exposure values within each pair. The rules for predicting the label for pair i is summarized in Table 4.1.

Table 4.1: Prediction of a New Pair Based on the Predicted Labels for Its Observed and Counterfactual Pairs

Label of the observed pair i	Label of the counterfactual pair $\tilde{i}$	Final label of the pair i
0	1	0
1	0	1
0	0	$label = \begin{cases} 0 & \Delta(i) > \Delta(\tilde{i}) \\ 1 & \Delta(i) \leq \Delta(\tilde{i}) \end{cases}$
1	1	$label = \begin{cases} 1 & \Delta(i) > \Delta(\tilde{i}) \\ 0 & \Delta(i) \leq \Delta(\tilde{i}) \end{cases}$

4.3.3 Effect Estimation

Matched Forest algorithm (MF) aims at identifying important exposure and matching variables from high-dimensional matched case-control data sets. Once important variables are identified, another step is to estimate the effect of important variables. In this section, we propose two metrics to measure the effect of important exposure variables identified by MF. These two metrics are suitable for variables with main effects. Thus, if multiple variables have been identified as important by MF, we first need to distinguish between variables with main effects and variables with interaction effects. One algorithm which can be used for this purpose is Iterative Random Forest (iRF) which was proposed by Basu *et al.* (2018) to detect variables with interactions.

Our method for estimation of effects is motivated by Conditional Logistic Regression (CLR) model, but we use a substantially different approach. CLR uses a linear logistic regression model and a conditional likelihood approach to estimate the coefficients associated with exposure variables. The coefficients estimated by CLR have the same interpretation as logistic regression. Each coefficient is interpreted as the change in logit for one unit increase in the corresponding exposure variable given all other variables are constant within each stratum. Our method defines two new metrics instead of logit function. These metrics are obtained based on the predicted labels and margin for observed case-control strata whose labels are equal to 0.

Let D represents our matched data set with N case-control strata labeled as 0, R exposure variables denoted by $\{x_1, x_2, \ldots, x_R\}$ and M matching variables denoted by $\{v_1, v_2, \ldots, v_M\}$. Let $(x_r^1(i), x_r^0(i))$ represents the value of exposure variable x_r for case and control units in stratum i respectively. Assume that we are interested in identifying the effect of exposure variable x_l. Let f denote MF algorithm trained

on matched data set D and $\hat{y}(i) \in \{0, 1\}$ denote the predicted label by MF for each case-control stratum i. The procedure to estimate $\hat{y}(i)$ was explained in Table 4.1. The first metric which we denote as M_1 is defined as

$$M_1 = \frac{1}{N} \sum_{i=1}^{N} \Delta(i) \mathbb{1}(\hat{y}(i) = 0) \tag{4.11}$$

where $\mathbb{1}(.)$ is the indicator function and $\Delta(i)$ is the margin for pair i computed based on Equation 4.10. The second metric which we denote as M_2 is defined as

$$M_2 = \frac{Pr_0}{Pr_1} \tag{4.12}$$

where $Pr_c = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}(i) = c)$ is the proportion of N strata classified as $c \in \{0, 1\}$ by MF.

Our method estimates the effect of exposure variable x_l as the change in metric M_1 or M_2 after $d \in \mathbb{R}$ $(d \neq 0)$ unit change in x_l for only case subjects, given that all other exposure and matching variables are left unchanged for each stratum. Formally, Let M_1^0 and M_2^0 denote metrics obtained by Equations 4.11 and 4.12 respectively using predictions from model f for matched data set D. Our method creates a new data set $\tilde{D}$ which is the same as D except for the exposure variable x_l. In matched data sets $\tilde{D}$, the case and control values of x_l are changed according to $(x_l^1(i) + d, x_l^0(i))$ for $i \in \{1, 2, \ldots, N\}$, where d is a nonzero scalar. Let M_1^d and M_2^d denote metrics obtained by Equations 4.11 and 4.12 respectively using predictions from model f for the new matched data set $\tilde{D}$. We denote the effect of exposure variable x_l for d unit change by $\phi^d(x_l)$ which is computed as

$$\phi^d(x_l) = M_1^d - M_1^0 \tag{4.13}$$

or

$$\phi^d(x_l) = M_2^d - M_2^0 \tag{4.14}$$

based on metric M_1 and M_2 respectively. A modified M_2 could also use the log of ratio in Equation 4.12.

4.4 Experiments

4.4.1 Simulations

We conducted simulation studies to demonstrate the effectiveness of proposed methods. In the first set of simulations, the objective is to compare WMF with MF and see how the performance of WMF changes depending on the number of iterations. In the second set of simulations, the objective is to evaluate our effect estimation method.

Simulation 1: WMF

The simulation designs implemented here are similar to Chapter 3. In particular, we generated matched pairs with no effect of exposure and 5 different effect types including linear, non-linear, interaction between a matching and an exposure, interaction between two exposures and interaction between three exposures. For each simulation, 100 data sets are generated. The number of matched pairs (N) in data sets generated for null scenario, linear and non-linear effects is 600, and it is increased to 800 for data sets generated with interaction effects. We also set the number of exposure variables (R) to 100 and number of matching variables (M) to 5. The values of all control variables x_r for $r \in \{1, 2, \ldots, R\}$, unless otherwise stated, are generated from a uniform distribution between 1 and 50 and case values are generated according to $x_r^0 + d_r$ where d_r follows normal distribution $N(\mu_r, 1)$. In our simulations, $\mu_r = 0$ for exposure values with no effect, $\mu_r \in \{1, 1.5, 2\}$ for exposure variables with positive effect and $\mu_r \in \{-0.5, -0.75, -1\}$ for exposure variables with negative effect. The

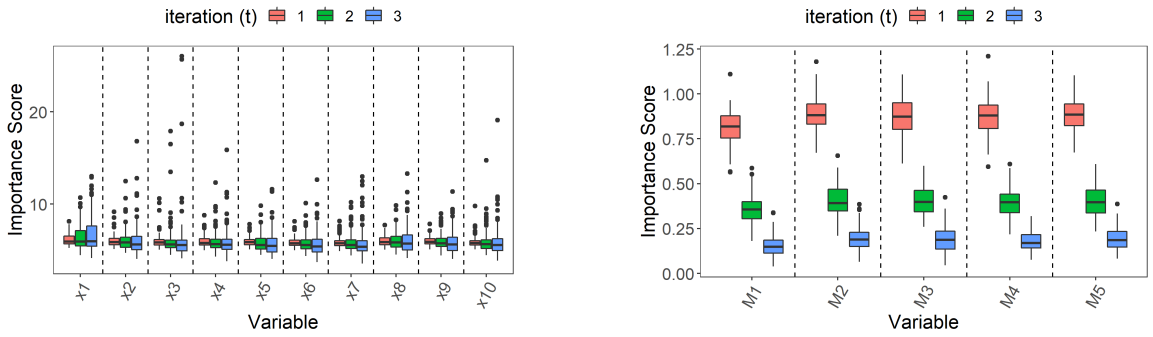
absolute value of μ_r ($|\mu_r|$) indicates the effect size. In addition, matching variables are generated from Poisson (5) distribution. For more details regarding how data sets are generated for each simulation, see Chapter 3.

We used the default parameters in R *randomForest* package at each iteration of WMF. In particular, we set number of trees to 500, number of variables selected at each split to $\sqrt{p}$ where p is the number of variables in the transformed data set (D^*) and grow trees to purity. The performance of WMF was tested at different number iterations $t \in \{1, 2, 3\}$. The WMF algorithm with $t = 1$ is equivalent to the MF algorithm. To select important variables from WMF, a similar approach to Chapter 3 was used, that is, observed MFI scores in each iteration were compared with the null distributions generated from randomly assigned labels in each pair. We compared the variable selection performance of WMF in three iterations using ROC curves which plots true positive rates (TPR) versus false positive rates (FPR) along different values of the significance level α. We also measured the variable selection accuracy of WMF as the area under the curve (AUC). For more details regarding the variable selection procedure, see Shomal Zadeh *et al.* (2020). For each simulation design, we show MFI scores of the top 20 exposure variables with the largest scores and all 5 matching variables at third iteration $k = 3$. Also, we show ROC curves at iterations $k \in \{1, 2, 3\}$ with their AUC.

Null scenario: Here, data sets are generated with no effect of exposure variables. Similar to Chapter 3, We used two generative models to simulate data sets with no effect of exposures. One experiment simulates data sets where exposure variable x_1 is associated with the matching variable v_1 and the other model simulates data sets where all variables are independent.

Figures 4.1 and 4.2 correspond to simulations where matching variable v_1 is associated with exposure variable x_1. From Figure 4.1, we can observe that MFI scores

of exposure and matching variables do not indicate any variable with significantly larger score than others and as t increases the MFI scores remain consistent or decrease slightly Figure 4.2 shows how WMF performs in selecting important variables. The false positive rates (FPR) are small at different values of α and do not change by increasing t. Therefore, WMF performs comparable to MF in terms of FPR and they both have small FPR. Figures 4.3 and 4.4 correspond to the simulations with independent variables. From Figure 4.3, we can observe that MFI scores of exposure and matching variables do not indicate any variable with significantly larger score than others and as t increases the MFI scores remain consistent or decrease slightly. Figure 4.4 shows how WMF performs in selecting important variables. The false positive rates (FPR) are small at different values of α and do not change by increasing t. Therefore, WMF performs comparable to MF in terms of FPR and they both have small FPR.



(a) MFI scores: exposure variables

(b) MFI scores: matching variables

Figure 4.1: Null Scenario Where Matching Variable v_1 Is Associated with Exposure Variable x_1: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iteration $t \in \{1, 2, 3\}$. MFI Scores of Exposure and Matching Variables Either Decrease Slightly or Remain Consistent with Increasing t.

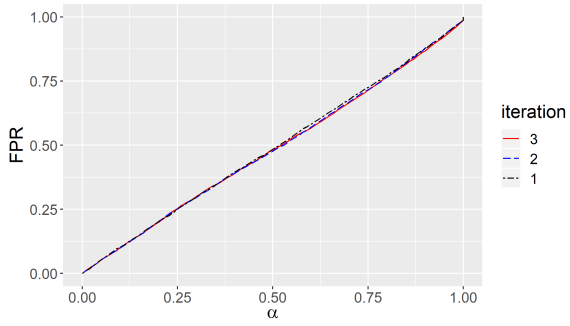
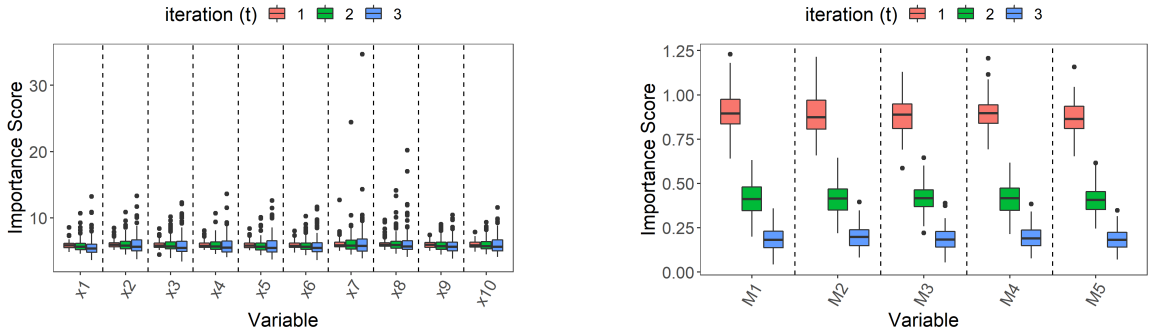


Figure 4.2: Null Scenario Where Matching Variable v_1 Is Associated with Exposure Variable x_1: Evaluating the Performance of WMF in Terms of Its FPR along Different Values of Significance Level α. FPR of WMF Remains Consistent by Increasing t.



(a) MFI scores: exposure variables

(b) MFI scores: matching variables

Figure 4.3: Null Scenario With Independent Variables: MFI Scores of (a) Exposure Variables And (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. MFI Scores of Exposure and Matching Variables Either Decrease Slightly or Remain Consistent with Increasing t.

Linear exposure effect: We simulated data sets where variable x_1 has a negative effect and all other variables are noise. To generate case values of x_1 for small, moderate, and strong effect size, μ_1 is set to $\{-0.5, -0.75, -1\}$ respectively for pairs with negative effect. The size of the effect is small relative to standard deviations. Figure 4.5 shows MFI scores of exposure and matching variables at $t \in \{1, 2, 3\}$ when

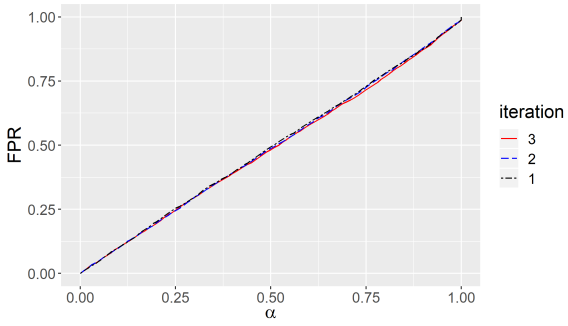
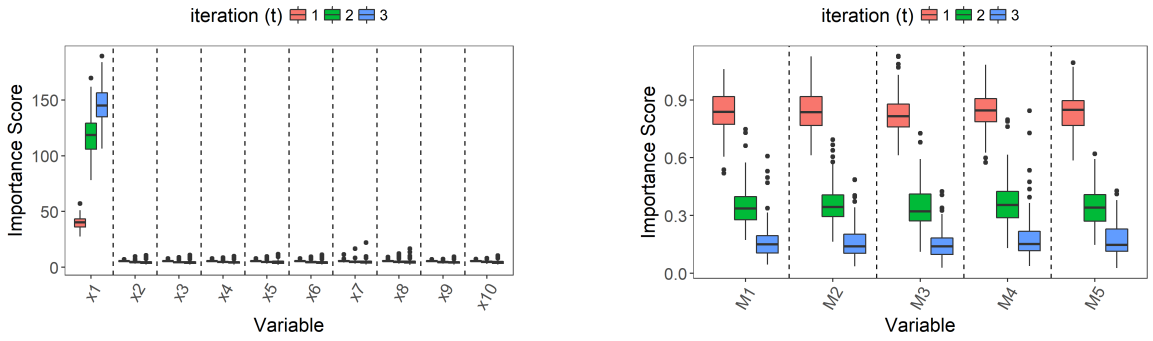


Figure 4.4: Null Scenario with Independent Variables: Evaluating the Performance of WMF in Terms of Its FPR along Different Values of Significance Level α. FPR of WMF Remains Consistent by Increasing t.

$\mu_1 = -1$ for pairs with negative effect. From Figure 4.5a, we observe that MFI score of exposure variable x_1 is larger that other exposure variables at all iterations and its MFI score increases with increasing t, while MFI scores of other exposure variables which are noise remain consistent as t increases. Figure 4.5b shows that MFI scores of matching variables decreases with increasing t due to their relatively small MFI scores compared to other variables. We also tested the variable selection accuracy of WMF at different levels of effect size. Figure 4.6 shows ROC curves and AUC for each level of effect size in iterations $t \in \{1, 2, 3\}$. We observe that WMF achieves the ideal variable selection accuracy (AUC=1) in all iterations. Therefore, WMF performs better than MF (WMF with one iteration) in identifying variables with linear effects and its performance improves as the number of iterations gets larger.

Non-linear exposure effect: Here, data sets are simulated with a non-linear effect for exposure variable x_1 and no other effect from other variables. Variable x_1 is generated such that there is a positive effect of x_1 when $x_1^0 < 25$ and a negative effect when $x_1^0 > 25$. To generate x_1^1 for small, moderate, and strong effect size, μ_1 is set to $\{-0.5, -0.75, -1\}$ for pairs with negative effect and $\{1, 1.5, 1\}$ for pairs with

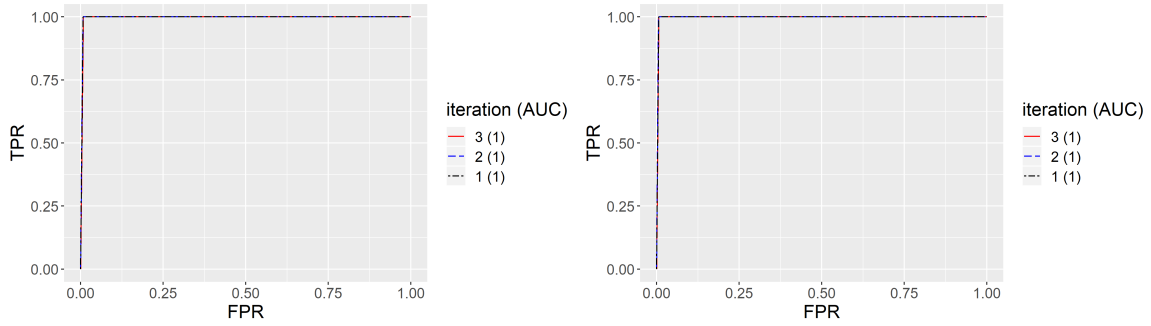


(a) MFI scores: exposure variables

(b) MFI scores: matching variables

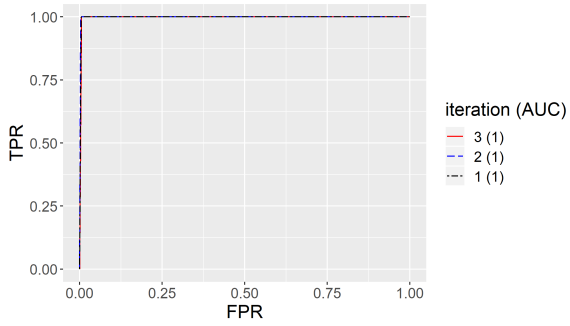
Figure 4.5: Linear Exposure Effect: MFI Scores of (a) Exposure Variables And (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with a Linear Effect of x_1 Where $\mu_1 = -1$ for Pairs with Negative Effect Of x_1. MFI Score of Exposure Variable x_1 Is Larger than Other Exposure Variables in All Iterations and It Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variables Have Small MFI Scores Which Drop with Increasing t.

positive effect. Figure 4.7 shows MFI scores of exposure and matching variables for simulation with $\mu_1 \in \{2, -1\}$. Figure 4.7a shows that MFI scores of exposure variable x_1 is significantly larger than other exposure variables at all iterations and it improves as t increases. However, the MFI scores of other exposure variables remain consistent with increasing t. From Figure 4.7b, there can be seen that MFI scores of matching variables drop by increasing t due to their relatively smaller MFI score at $t = 1$ compared to other variables in data sets. We also evaluated the performance of WMF at different levels of effect size. Figure 4.8 shows ROC curves and AUC for each level of effect size in iterations $t \in \{1, 2, 3\}$. The ROC curve and AUC are comparable or better for increasing t. As t increases, the amount of improvement is larger when the effect size is small ($\mu_1 \in \{1, -0.5\}$). Therefore, WMF performs better



(a) Small effect ($\mu_1 = -0.5$)

(b) Moderate effect ($\mu_1 = -0.75$)

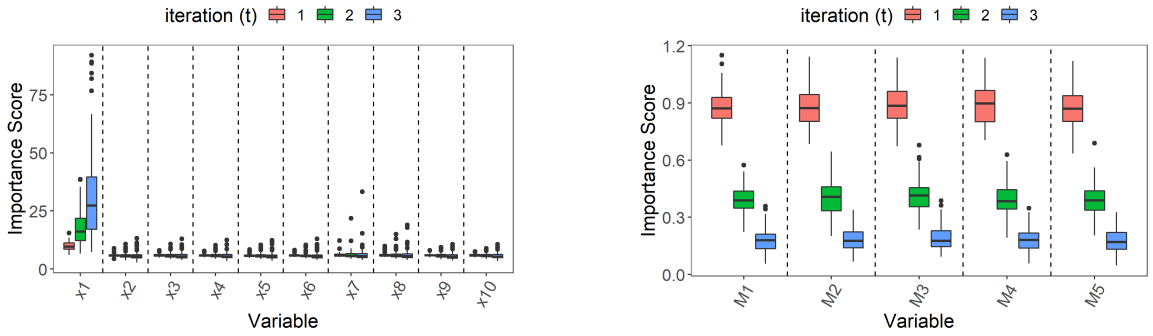


(c) Strong effect ($\mu_1 = -1$)

Figure 4.6: Linear Exposure Effect: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Is 1 for All Three Effect Size in Different Iterations.

than MF (WMF with one iteration) in identifying variables with non-linear effects and its performance improves as the number of iterations gets larger, especially when effect size is smaller.

Interaction between a matching and an exposure: Here data sets are simulated with an interaction effect between matching variable v_1 and exposure variable x_1. The data generative model in this simulation study is similar to Shomalzadeh *et al.* (2019). Variables v_1, x_1^0 and x_1^1 are generated so that the exposure variable x_1 has positive effect for smaller values of v_1 and negative effect for larger values of v_1. To generate pairs with negative and positive effects, μ_1 is set to $\{-0.5, -0.75, -1\}$



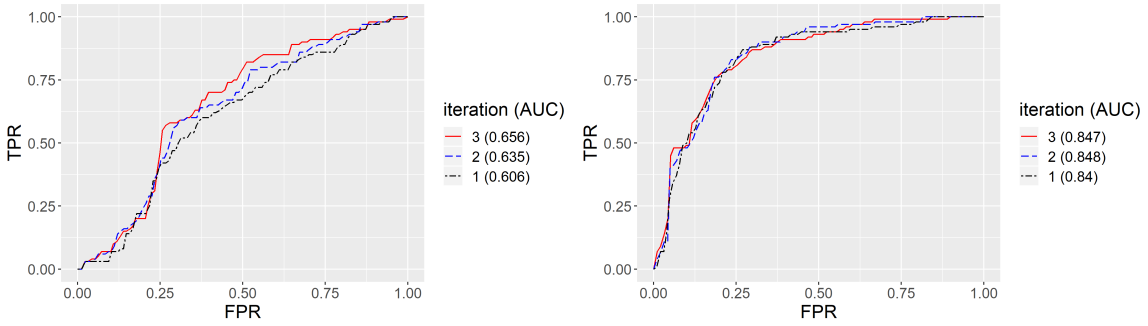
(a) MFI scores: exposure variables

(b) MFI scores: matching variables

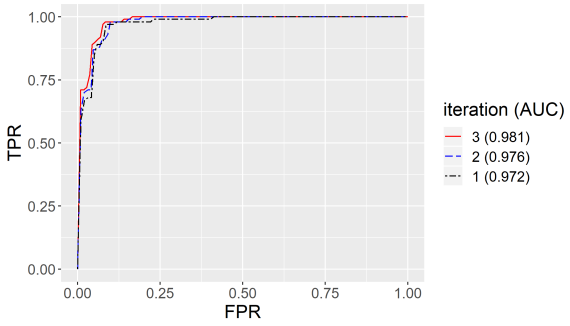
Figure 4.7: Non-linear Exposure Effect: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with a Non-linear Effect of x_1 Where $\mu_1 \in \{-1, 2\}$. MFI Score of Exposure Variable x_1 Is Larger than Other Exposure Variables in All Iterations and It Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variables Have Small MFI Scores Which Drop with Increasing t.

and $\{1, 1.5, 2\}$, respectively. For a small effect, $\mu_1 \in \{-0.5, 1\}$, a moderate effect, $\mu_1 \in \{-0.75, 1.5\}$ and for a strong effect, $\mu_1 \in \{-1, 2\}$.

Figure 4.9 shows MFI scores of exposure and matching variables when $\mu_1 \in \{-1, 2\}$. MFI score of exposure variable x_1 is significantly larger than other exposure variables and its MFI score improves with increasing the number of iterations. Other exposure variables which are noise have relatively smaller MFI scores than exposure variable x_1 and their scores remain consistent with increasing t. Similarly, matching variable v_1 has a significantly larger MFI score than other matching variables and its score increases with t, while other matching variables have a relatively smaller MFI score which is consistent with increasing t. We also tested the performance of WMF in selecting important variables at different effect size. Figure 4.10 shows ROC



(a) Small effect $\mu_1 \in (1, -0.5)$ (b) Moderate effect $\mu_1 \in (1.5, -0.75)$

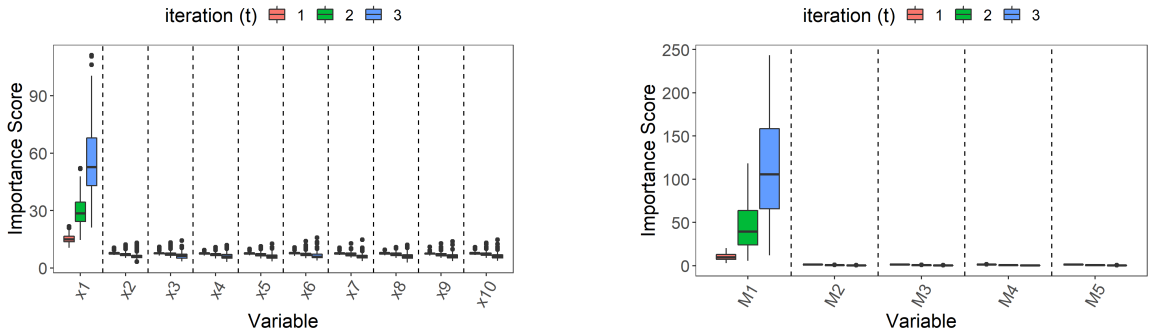


(c) Strong effect $\mu_1 \in (2, -1)$

Figure 4.8: Non-linear Exposure Effect: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves or Remains Consistent with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small.

curves and their AUC for small, moderate, and strong effects. There can be seen that the performance of WMF improves by increasing t. Also, the amount of increase in AUC is larger when the effect size is small. Therefore, WMF performs better than MF (WMF with one iteration) in identifying interaction effects between matching and exposure variables and its performance improves as the number of iterations gets larger, especially when effect size is smaller.

Interaction between two exposures: data sets are generated with a two-way interaction between exposure variables x_1 and x_2 according to Shomal Zadeh *et al.*

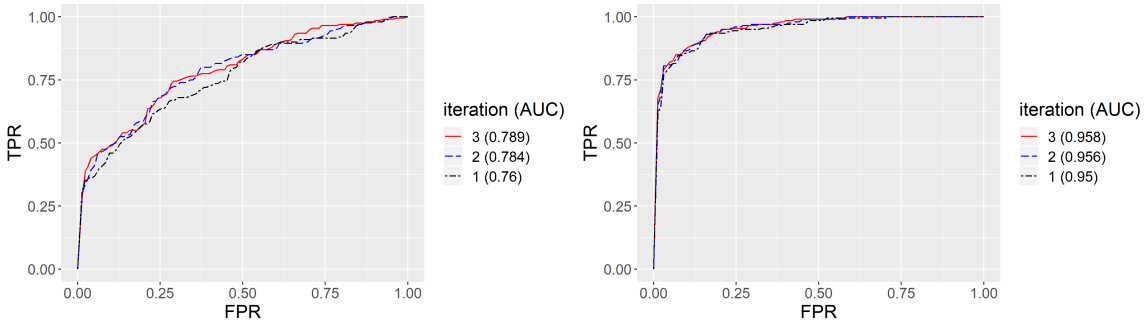


(a) MFI scores: exposure variables

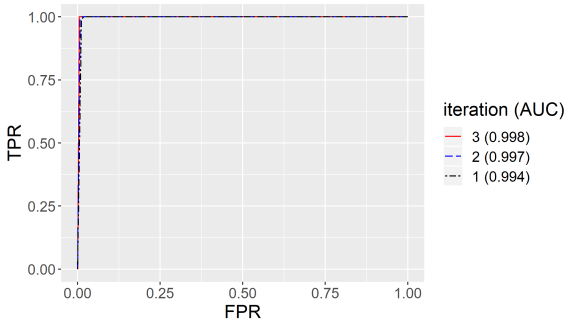
(b) MFI scores: matching variables

Figure 4.9: Interaction Effect Between a Matching and An Exposure: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with An Interaction Effect Between Matching Variable v_1 and Exposure Variable x_1 Where $\mu_1 \in \{-1, 2\}$. MFI Score of Exposure Variable x_1 Is Larger than Other Exposure Variables in All Iterations and It Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variable v_1 Has Larger MFI Score than Other Matching Variables and It Improves with Increasing t, While the MFI Scores of Other Matching Variables Remain Consistent with Increasing t.

(2020) such that the effect of variable x_2 changes at different levels of x_1. To generate x_1^1 for pairs with negative and positive effects, $\mu_1 \in \{-1, 2\}$ and to generate x_2^1 for pairs with negative and positive effect, $\mu_2 \in \{-0.5, 1\}$ for a weak effect, $\mu_2 \in \{-0.75, 1.5\}$ for a moderate effect, and $\mu_2 \in \{-1, 2\}$ for a strong effect. Figure 4.11 shows MFI scores of exposure and matching variables for $\mu_2 \in \{-1, 2\}$ in iterations $t \in \{1, 2, 3\}$. Figure 4.11a shows that MFI scores of exposure variables x_1 and x_2 are larger than other exposure variables and their scores improve with increasing t. However, MFI scores of other exposure variables which are noise remain consistent or decrease with increasing t. From Figure 4.11b, we observe that MFI scores of matching variables



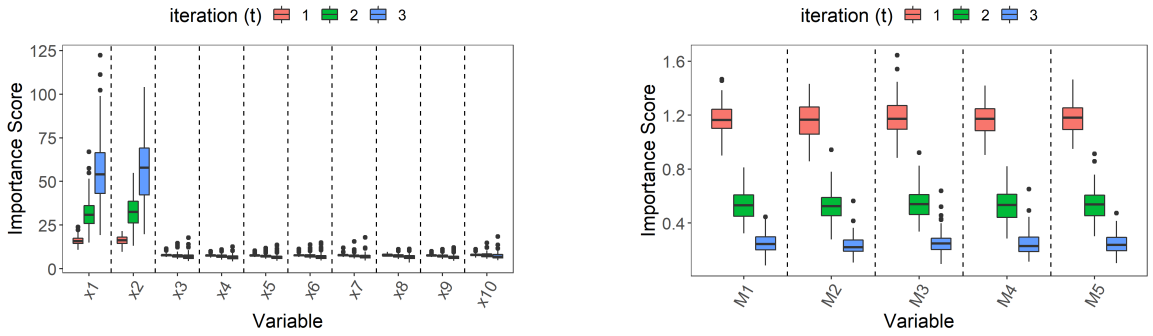
(a) Small effect $\mu_1 \in \{1, -0.5\}$ (b) Moderate effect $\mu_1 \in \{1.5, -0.75\}$



(c) Strong effect $\mu_1 \in \{2, -1\}$

Figure 4.10: Interaction Effect Between a Matching and An Exposure: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small.

drop as number of iterations increases from 1 to 3. We also evaluated the performance of WMF at different levels of effect size. Figure 4.12 shows ROC curves and AUC of WMF for data sets simulated with weak, moderate, and strong effect in iterations $t \in \{1, 2, 3\}$. ROC curve and AUC improve or remain consistent as t increases and the amount of improvement is larger when the effect size is small. Therefore, WMF performs better than MF (WMF with one iteration) in identifying interaction effects between two exposure variables and its performance improves as the number of iterations gets larger, especially when effect size is smaller.

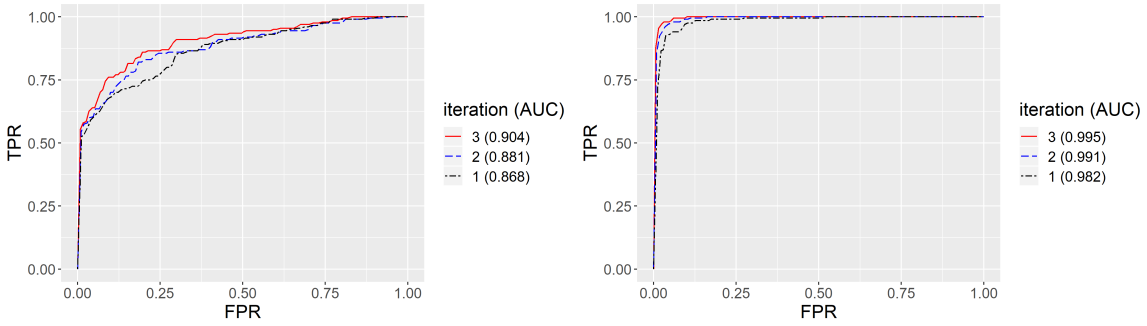


(a) MFI scores: exposure variables

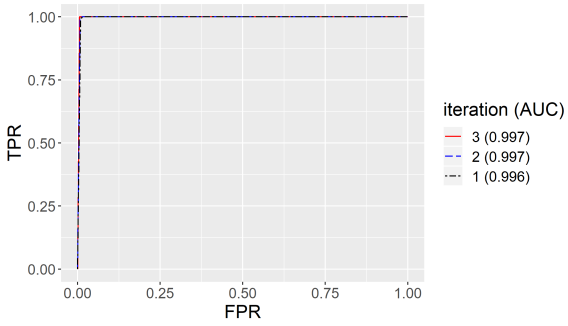
(b) MFI scores: matching variables

Figure 4.11: Interaction Effect Between Two Exposure Variables: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with An Interaction Between Exposure Variables x_1 and x_2 Where $\mu_1 = \mu_2 \in \{-1, 2\}$. MFI Score of Exposure Variables x_1 and x_2 Are Larger than Other Exposure Variables in All Iterations and Their Scores Improve with Increasing t, While the MFI Scores of Other Exposure and Matching Variables Remain Consistent or Decrease with Increasing t.

Interaction between three exposures: Here, we generated an interaction effect between exposure variables x_1, x_2, and x_3 and no effect from other variables. That is, data sets are generated where the effect of x_3 changes at different levels of x_1 and x_2. To generate x_1^1 and x_2^1, $\mu_1 = \mu_2 = -1$ for pairs with negative effect and $\mu_1 = \mu_2 = 2$ for pairs with positive effect. Also, to generate variable x_3^1 for pairs with negative and positive effects, $\mu_3 \in \{-0.5, 1\}$ for a small effect, $\mu_3 \in \{-0.75, 1.5\}$ for a moderate effect, and $\mu_3 \in \{-1, 2\}$ for a strong effect. Figure 4.13 shows MFI scores of exposure and matching variables when $\mu_3 \in \{-1, 2\}$. From Figure 4.13a, we can observe that variables x_1, x_2 and x_3 received relatively larger MFI scores than other exposure variables in all three iterations and their scores improve as t increases. However, MFI scores of other exposure variables which are noise remain



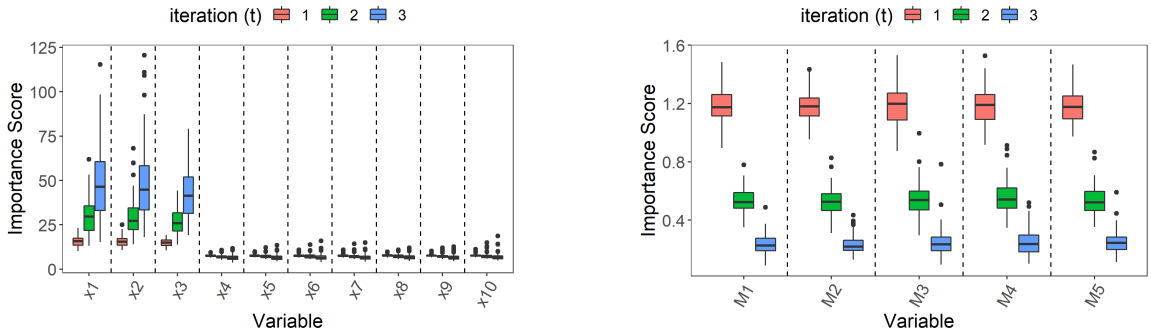
(a) Small effect $\mu_2 \in \{1, -0.5\}$ (b) Moderate effect $\mu_2 \in \{1.5, -0.75\}$



(c) Strong effect $\mu_2 \in \{2, -1\}$

Figure 4.12: Interaction Effect Between Two Exposure Variables: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small.

consistent for increasing t. Also, Figure 4.13b shows that MFI scores of matching variables drop with increasing t. We also tested the performance of WMF in detecting effects with different strengths, whose results are shown in Figure 4.14. It can be seen in Figure 4.14 that ROC curve and AUC improve as t increases and larger improvement is achieved when the effect size is small. Therefore, WMF performs better than MF (WMF with one iteration) in identifying interaction effects between three exposure variables and its performance improves as the number of iterations gets larger, especially when effect size is smaller.

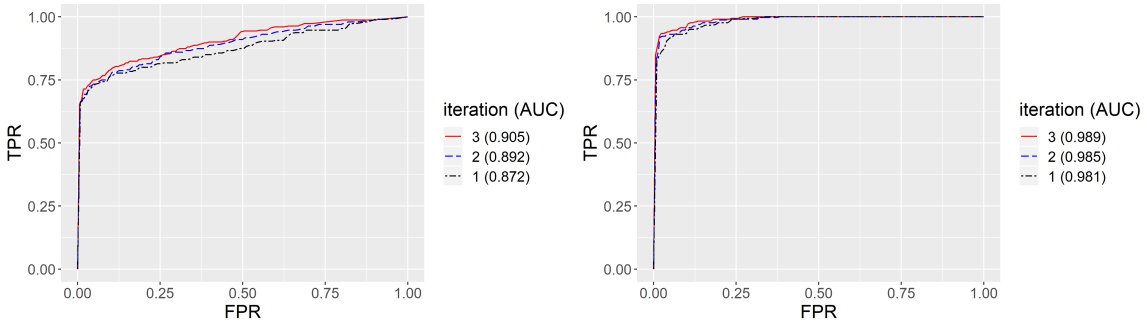


(a) MFI scores: exposure variables

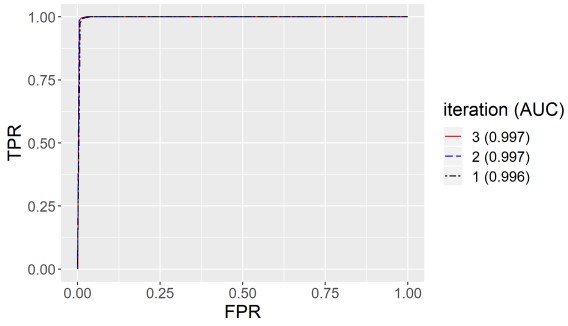
(b) MFI scores: matching variables

Figure 4.13: Interaction Effect Between Three Exposure Variables: MFI Scores of (a) Exposure Variables and (b) Matching Variables in Iterations $t \in \{1, 2, 3\}$. Results Are Shown for Data Sets Simulated with An Interaction Between Exposure Variables x_1, x_2 and x_3 Where $\mu_1(-) = \mu_2(-) = \mu_3(-) = -1$ and $\mu_1(+) = \mu_2(+) = \mu_3(+) = 2$. MFI Score of Exposure Variables x_1, x_2 and x_3 Are Larger than Other Exposure Variables in All Iterations and Their Score Improves with Increasing t, While the MFI Scores of Other Exposure Variables Remain Consistent with Increasing t. Matching Variables Have Small MFI Scores Which Drop with Increasing t.

Summary of simulation results: We observed in our simulations that WMF performs better than MF in identifying variables with different effect types including linear, non-linear, interactions between exposure variables, and interactions between matching and exposure variables. We showed in our experiments in Chapter 3 that MF is better than competing algorithms including CLR and the boosting method (Adewale *et al.* (2010)), thus, WMF also outperforms these algorithms. We also observed in our simulations as the number of iterations increases in WMF, its variable selection improves, and the amount of improvement is larger when effect size is smaller.



(a) Small effect $\mu_3 \in \{1, -0.5\}$ (b) Moderate effect $\mu_3 \in \{1.5, -0.75\}$



(c) Strong effect $\mu_3 \in \{2, -1\}$

Figure 4.14: Interaction Effect Between Three Exposure Variables: Evaluating WMF Performance by Effect Size Using ROC Curves. The Variable Selection Accuracy (AUC) Improves with Increasing t and the Amount of Improvement Is Larger When the Effect Size Is Small.

Simulation 2: Effect Estimation

Here, we simulated a matched data set with linear exposure effect according to the simulation design described in Section 4.4.1. In particular, we simulated a matched data set with 600 strata, 100 exposure variables, and 5 matching variables. The exposure variable x_1 is simulated with a negative effect where for all strata μ_1 is set to -1, and the remaining exposure variables $\{x_2, x_3, \ldots, x_{100}\}$ are simulated with no effect ($\mu_r = 0$).

To estimate the effect of exposure variable x_1, we used Neural Network as the classifier instead of Random Forest algorithm. We selected Neural Network because it is more sensitive to the change in feature values than Random Forest algorithm. The Neural Network model used in this simulation study consists of 3 hidden layers of size 10 with Relu activation function followed by a softmax output layer with 2 nodes. The number of epochs is set to 30, batch size to 5 and learning rate to 0.001. Also, $L1$ regularization with parameter 0.007 is used for the weights connecting input layer to the first hidden layer to handle overfitting. Using other hyper parameter values and Neural network architectures might improve the results.

We used both metrics M_1 and M_2 to evaluate the effect of exposure variable x_1 for $d \in \{-1.5, -1, -0.5, 0.5, 1, 1.5\}$. To estimate the classification probabilities, we used $5-$fold cross validation to split strata into training and test sets. The detailed procedure is summarized as follows:

1. The matched case-control pairs are randomly split into 5 disjoint sets $s = 1, 2, \ldots, 5$. Let $D(s)$ shows feature values in the transformed data set corresponding to matched pairs in set s.

2. For each set $s \in \{1, 2, \ldots, 5\}$:

   (a) Train Neural Network on the transformed data set including matched pairs in all sets except s.

   (b) Predict classification probabilities for matched sets in test set s with feature values $D(s)$.

   (c) Add d units to the value of exposure variable x_1 for case units in test set s. Let $D'(s)$ denote new feature values in the transformed data set corresponding to matched pairs in set s.

(d) Predict classification probabilities for matched sets in test set s with feature values $D'(s)$.

3. Compute $\phi^d(x_1)$ based on metrics M_1 and M_2 using Equations 4.11 and 4.12 respectively.

Figure 4.15 shows the results of simulation. The measure of effect size based on metric M_1 is shown in red color (triangle) on the right y-axis, and the measure of effect size based on M_2 is shown in blue color (circle) on the left y-axis. Each point on this plot shows effect size corresponding to a metric (M_1 or M_2) and a value of d. We observe in Figure 4.15 that when d is set to a positive value, both measures of effect size are negative. This shows that our method correctly identifies the sign of effect. We also observe larger variation in the measures of effect size based on M_1 when $d > 0$ than $d < 0$. However, an opposite pattern is seen for effect size measures based on M_2. That is, effect size is more sensitive to the change in d for $d < 0$ than $d > 0$.

4.4.2 Case Studies

Our case studies include 2 biomedical data sets, namely, Statlog heart disease (Lichman (2013)) and childhood acute lymphoblastic leukemia study (Bhojwani *et al.* (2006)). The childhood acute lymphoblastic leukemia study consists of matched case-control pairs. The Statlog heart disease data set does not have a matched design, so it is first converted into matched pairs before analysis. We used R *MatchIt* package for the exact matching of controls to cases. Further details regarding the matching procedure have been provided in the corresponding section.

We evaluated the performance of proposed methods using these case studies. We set number of iterations in WMF to 3 and use the default parameters of R *random-*

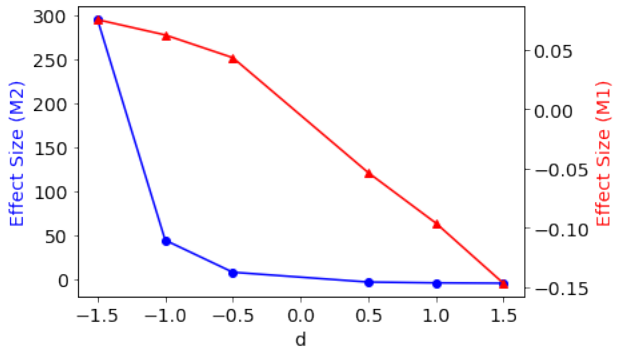


Figure 4.15: Effect Size Measures Based on Metrics M_1 and M_2 for a Variable Simulated with Negative Linear Effect. The Effect Size Is Estimated for $d \in \{-1.5, -1, -0.5, .5, 1, 1.5\}$. The Measure of Effect Size Based on M_1 Is Shown in Red (Triangles) on the Right Y-axis and the Measure Based on M_2 Is Shown in Blue (Circle) on the Left Y-axis.

Forest package ($ntree = 500$, $mtry = \sqrt{p}$ and $mxnodes = NULL$) to run WMF, except for the childhood acute lymphoblastic luekemia study which was evaluated with $50,000$ decision trees due to its large number of variables. Also, to account for randomness in WMF, each algorithm was run in 100 replicates on each data set.

For variable selection, we used a method similar to Shomal Zadeh *et al.* (2020). To generate null distribution for each variable's MFI score, we permuted the case and control instances within each stratum 100 times and ran WMF on each permuted data set once. The generated null distributions were compared with the original MFI scores to select variables with MFI scores significantly large at a predetermined significance level α.

We evaluated the variable selection performance of WMF and classification accuracy of MF (WMF with one iteration) on both case studies. The classification accuracy of MF is compared with CLR for statlog heart disease data set. However, ClR does not converge for the childhood acute lymphoblastic luekemia study due to

its large number of variables, thus, WL_2Boost is used for comparison on this data set. We used 10−fold cross validation and ran both MF and CLR 100 times to account for randomness in cross-validation. The average of accuracy over 100 runs of MF and CLR was used as the evaluation metric.

Case Study 1: Statlog Heart Disease Data Set

This data set includes 120 subjects with heart disease (case) and 150 subjects without heart disease (control). The subjets in this data set are not matched, thus, we created matched pairs by matching each case with a control using variables age (discretized by 5-year intervals) and gender. The resulting matched data set has 80 case-control pairs. We analyzed the effect of 6 numerical exposure variables on heart disease. These exposures include Resting Blood Pressure (x_1), Serum Cholestoral (x_2), max heart rate (x_3), Oldpeak (x_4), Slope of peak ST segment (x_5), and Major vessels colored (x_6).

Figure 4.16 and Figure 4.17 show respectively MFI scores and p-values of WMF in iterations $t \in \{1, 2, 3\}$. Variables x_3, x_4, x_6 received relatively large MFI scores compared to other exposure variables and their MFI scores increase as t increases. The WMF algorithm also selects variables x_5 and x_6 as important in all iterations and at significance level $\alpha = 0.05$ (Figure 4.17). By examining the scatterplots of control versus case for each exposure variable, we observed that variables x_3, x_4, x_5, and x_6 potentially have an effect. In particular, for these 4 variables, we observed larger difference between the number of pairs with case greater than control and pairs with control greater than case. The relatively larger MFI scores of x_3, x_4 and x_6 also indicate that they are important. Although the p-values corresponding to x_3 and x_4 decreased by an increase in the number of iterations in WMF, they are still not significant at $\alpha = 0.05$. Thus, further modification in variable selection method of

WMF is required to improve its power. The reason why variable x_5 did not receive a large MFI score is that it only has 3 unique values and Strobl *et al.* (2007) showed in their experiments that RF's Gini importance score is smaller for variables with small number of unique values. Strobl *et al.* (2007) suggested using RF's permutation score when variables differ in the number of unique values. The variables selected by MF were compared with CLR in Shomal Zadeh *et al.* (2020). CLR selected variables x_3 and x_6 as important at $\alpha = 0.05$. Using all variables in data set, the classification accuracy of MF was compared with CLR. The average of accuracy over 100 runs of each algorithm was used as the metric for comparison. MF achieves a classification accuracy of 80% which is comparable with the classification accuracy of CLR (79%).

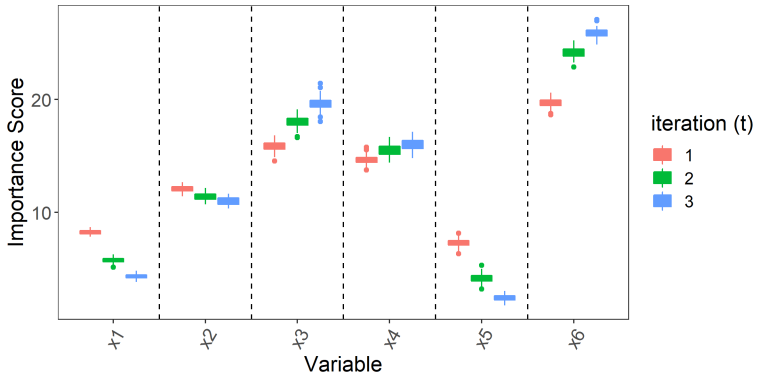


Figure 4.16: Statlog Heart Disease Data Set: MFI Scores of WMF at $t \in \{1, 2, 3\}$. Variables x_3, x_4 and x_6 Received Relatively Larger Scores than Other Exposure Variables and Their MFI Scores Improve as t Increases.

Case study 2: Childhood Acute Lymphoblastic Leukemia Study

The childhood acute lymphoblastic leukemia study is a matched pair study design conducted by Bhojwani *et al.* (2006) to identify factors leading to relapse. They selected 35 children with acute lymphoblastic leukemia who were relapsed after therapy

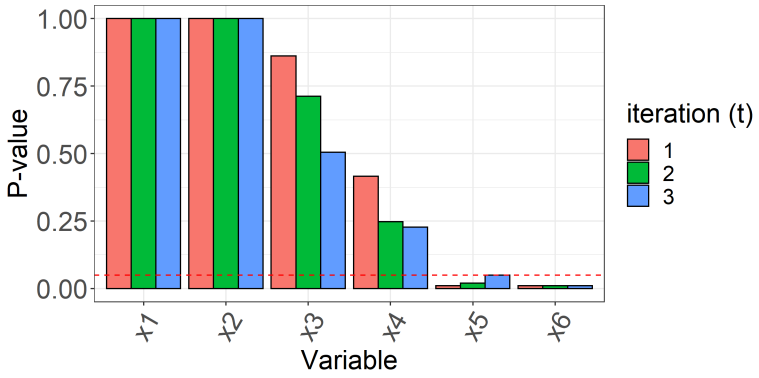


Figure 4.17: Statlog Heart Disease Data Set: P-values Computed by WMF at $t \in \{1, 2, 3\}$. Variables x_5 and x_6 Were Selected by WMF in $t \in \{1, 2, 3\}$ at Significance Level $\alpha = 0.05$.

and analyzed $22,283$ gene expression profiles in bone marrow of diagnosis (control) and relapsed (case) samples taken from the same patient. Thus, this data set contains 35 matched pairs and $22,283$ exposure variables.

We ran WMF on this data set with 100 replicates and $50,000$ trees. The average of MFI scores in each iteration were ranked from 1 (the most important) to $22,283$ (the least important) and important variables were selected at $\alpha = 0.01$. Figure 4.18 shows MFI scores in the third iteration of WMF for the top 50 variables with highest MFI scores. The 50 most important variables selected in iteration three were ranked no lower than 82 and 68 by the first and second iterations of WMF, respectively. Using the variable selection method based on permuting the labels within each pair, WMF selects 283, 144 and 177 variables in first, second, and third iterations, respectively. Thus, the regularization in WMF leads to smaller number of variables to be selected than MF (WMF at $t = 1$). We also evaluated the classification accuracy of WMF using the 50 most important variables identified in the last iteration of WMF. Using 10-fold cross-validation, MF achieves a classification accuracy of 98%,

which is substantially higher than 77% classification accuracy from WL_2Boost by using its 12 selected variables and 5-fold cross validation.

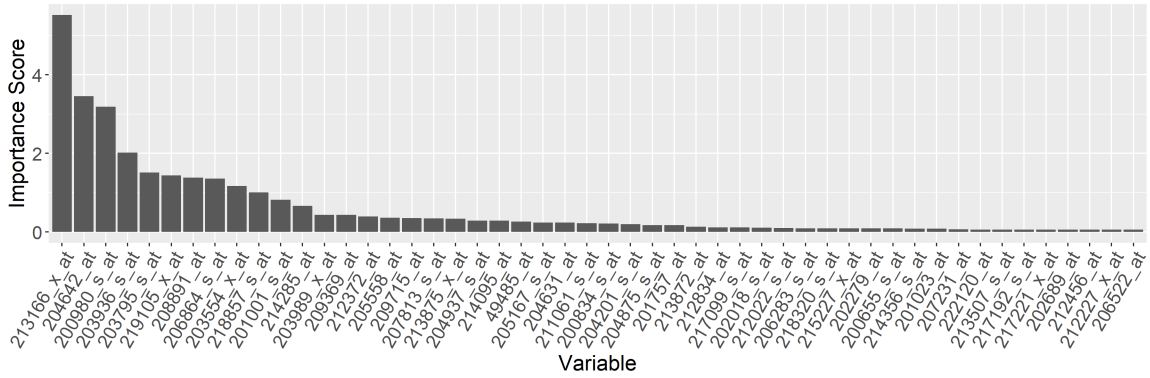


Figure 4.18: Childhood Acute Lymphoblastic Leukemia Study: MFI Scores of Top 50 Important Variable from the Third Iteration of WMF.

4.5 Conclusion

We presented three enhancements of MF, a powerful variable selection algorithm for high-dimensional matched case-control data sets. In particular, we proposed a regularized version of MF with improved power, a prediction algorithm based on MF to classify matched pairs into case-control or control-case, and two interpretable measures to estimate the effect of important variables identified by MF.

We tested the performance of proposed methods through extensive simulations and case studies. Our results demonstrate the effectiveness of WMF in selecting important variables. WMF is more useful for high-dimensional data sets with few informative variables and small effect size. We also observed in our case studies that MF has better classification accuracy than its competing algorithms for high-dimensional matched case-control data sets. The two metrics that we proposed for effect estimation can be effectively used to measure the effect of important variables identified by MF.

Chapter 5

MATCHED CASE-CONTROL ANALYSIS USING NEURAL NETWORKS

Matched case-control study designs are commonly used in healthcare, social, and behavioral sciences due to their ability to remove the effect of confounding and improve the efficiency in identifying important features. In matched case-control studies, observations with the outcome of interest (case) are grouped with observations without the outcome of interest (control) based on some matching variables. Often the goal of matched case-control studies is to identify matching and exposure variables informative in distinguishing between case and control units. The number of cases and controls within each matched set or stratum can vary, but usually one case is matched with a fixed number of controls that ranges between 1 to 5 (Hosmer Jr *et al.* (2013)). This study design is referred as matched $1 - L$ case-control, where L is the number of controls matched to each case. In the special case where $L = 1$, the study design is referred as matched pairs too. The common reason why more than one control is matched to a case or treated unit in matched case-control studies is to increase the power of a test of no effect of exposure variables assuming that matching variables are sufficient to remove the bias from nonrandom treatment assignment (Rosenbaum (2013)). In healthcare applications, matching is usually done on demographic variables such as age and gender to remove the effect of confounding variables which can lead to spurious results. For example, people with a disease may be matched with people without a disease based on their age and sex to identify gene expression markers significantly different between case and control observations. Regular machine learning algorithms such as Random Forest and Neural Network cannot be directly applied to matched case-control data sets because they assume observations

in data are independent. However, in matched case-control studies, correlation exists among units within a matched set, thus, specific methods are required to account for the matching structure of data. Analysis of matched case-control study designs will become challenging when data sets are high-dimensional with hundreds and thousands of variables or when the relationship between variables and outcome is highly nonlinear and involves interaction.

To address these challenges which exist in many of the modern real data sets, complex models such as neural networks are required. Neural networks have achieved high accuracy in several applications, e.g computer vision, that involve hundreds or thousands of variables. Also, recent innovations have enabled users to explain the predictions of neural networks in terms of input variables. One of the most recent explanation methods for neural networks is DeepSHAP (Lundberg and Lee (2017) and Chen *et al.* (2019)) that estimates Shapely values (Shapley (1953)) for neural networks. These properties of neural networks motivated us to adapt neural networks to matched case-control study designs.

Here, we present a neural network based approach which we call matched neural network (MNN), to assign importance scores to variables in a high-dimensional matched case-control study design. We first explain our method for matched pairs and then extend it to matched $1 - L$ case-control study designs with $L > 1$. Our method first transforms data to a supervised setting, then, a neural network classifier is trained on this data and importance of each variable is computed using DeepSHAP. The data transformation method is motivated by Shomal Zadeh *et al.* (2020) and SHAP scores computed by the neural network model are modified for the matched study.

Section 5.1 explains existing variable selection methods for high-dimensional matched case-control studies and provides a description of DeepSHAP method for

assigning importance scores to variables based on the predictions of neural network models. In section 5.2, we explain our method, matched neural network. Section 5.3 shows the effectiveness of our method through simulation and case studies. Section 5.5 concludes this research work.

5.1 Background

5.1.1 Variable Selection Methods For High-dimensional Matched Case-control Datasets

Majority of algorithms proposed for high-dimensional matched case-control data sets are based on the conditional logistic regression (CLR) model, for example Balasubramanian *et al.* (2014), Asafu-Adjei *et al.* (2017), and Qian *et al.* (2014) all use a CLR-based approach to identify important variables from matched case-control studies. They use a linear model with conditional likelihood function which is supplemented with cross products of two or more variables to handle interactions. However, these methods could become intractable in high-dimensional settings with hundreds or thousands of variables. There are also other algorithms which are not based on CLR. For example, Adewale *et al.* (2010) proposed two variants of boosting algorithm for data sets with correlated binary outcome. The first method is based on gradient descent boosting algorithm that uses a modified loss function to handle correlations among data. The second method modifies the likelihood optimization boosting algorithm using a generalized linear mixed model. The drawback of these methods could be their poor performance to detect variables with non-linear effects or interaction effects with other variables. Also, Stanfill *et al.* (2019) proposed a data preprocessing approach to generalize classification algorithms to matched case-control data sets, and referred to these modified methods as Conditional Classification algorithms. Their method centers each strata by the mean values of exposure and map the exposure

value of each unit to its difference from the center. This method does not handle dependency among units within a stratum, which is recommended by statistical principles. It breaks each stratum into multiple instances which are known to be dependent. Our experiments show that this method has difficulty detecting variables with non-linear and interaction effects. Matched Forest (MF) proposed by Shomal Zadeh *et al.* (2020) is also a recent method for variable selection from high-dimensional matched case-control data sets. This method transforms a matched case-control data set to a supervised setting which accounts for the matching structure of data and then applies Random Forest (RF) on the transformed data set to compute variable importance score and select important variables. This method is able to detect nonlinear effects and interactions between exposure and matching variables in high-dimensional matched case-control data sets.

5.1.2 DeepSHAP

DeepSHAP is a feature attribution method designed for deep neural networks. It modifies DeepLIFT algorithm (Shrikumar *et al.* (2017) and Shrikumar *et al.* (2016)) to estimate SHAP (SHapley Additive exPlanations) values (Lundberg and Lee (2017)) over the feature space for each individual prediction. In what follows, we first explain the general approach for computing SHAP values, and then explain how SHAP values are estimated by DeepSHAP through a modification on DeepLift attribution algorithm.

SHAP is connected with Shapley values (Shapley (1953)) from game theory which explains the output of any machine learning model $y = f(x)$ by assigning an importance value to each feature x_j ($\phi(x_j)$) that represents the effect of including that feature on prediction. To compute this effect for feature x_j, a model $f_{S\cup\{j\}}(x_{S\cup\{j\}})$ is trained on a subset of features $S \subseteq F$, where F is the set of all features, with feature

x_j present and another model $f_S(x_S)$ is trained on the feature subset S with feature x_j withheld. The marginal effect of feature x_j when it is added to feature subset S is then computed by $f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S)$. When the model f is non-linear or feature variables are not independent, the marginal contribution of a variable x_j depends on the other features in the model (S), thus, shapley values arise from the weighted average of marginal contributions over all possible feature subsets $S \subseteq F \setminus \{j\}$:

$$\phi(x_j) = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}[f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S)] \tag{5.1}$$

where $|S|$ represents the size of set S. For many of the machine learning models, it is not feasible to predict the output of the model for a subset of features. SHAP (Lundberg and Lee (2017)) uses a conditional expectation function of the original model to define simplified input features. That is, it defines $f_x(S)$ by $E[f(x)|x_S]$ which is the expected value of the model conditioned on the feature subset S. The exact computation of SHAP values is challenging for complex models. However, existing additive feature attribution models can be modified to approximate SHAP values.

DeepLift is an additive feature attribution method for neural network which is modified in DeepSHAP to approximate SHAP values. DeepLift is one of backpropagation-based approaches that propagates an importance signal from an output neuron through hidden layers and finally to input features. This is computationally efficient because importance scores are computed in only one backward pass. DeepLIFT assigns to each feature x_j a contribution score $C_{\Delta x_j \Delta y}$ that represents the amount of difference in output y from a reference y^0 attributed to the difference of that feature x_j from the reference x_j^0. The choice of a reference depends on domain-specific knowledge. For example, Shrikumar *et al.* (2017) uses an image with all zeros as the reference for MNIST data set because this is the background

of all images in this data. DeepLIFT is an additive feature attribution method that follows "summation-to-delta" property:

$$\sum_{j=1}^{n} C_{\Delta x_j \Delta y} = \Delta y \tag{5.2}$$

where $\Delta y = f(x) - f(x^0)$ and $\Delta x_j = x_j - x_j^0$. Similar to how chain rule is constructed for partial derivatives to compute the gradient of the output with respect to an input, DeepLIFT uses "chain rule for multipliers" to compute the global multiplier for any neuron to a given target neuron via backpropagation. For a given input x_j and target neuron y, the multiplier is defined as:

$$m_{\Delta x_j \Delta y} = \frac{C_{\Delta x_j \Delta y}}{\Delta x_j} \tag{5.3}$$

which is the contribution of input x_j to target neuron y divided by the difference-from-reference of the input Δx_j. According to the chain rule for multipliers, the global multiplier from input x_j to target neuron y ($m_{\Delta x_j \Delta y}$) is computed by recursively passing the multipliers backward through the network and summing them up over all paths connecting input x_j to target neuron y. Assuming that there is a hidden layer with neurons $h_1, \ldots, h_n$ between input neuron x_j and target neuron y, the global multiplier for x_j to y is computed as follows:

$$m_{\Delta x_j \Delta y} = \sum_i m_{\Delta x_j \Delta h_i} \times m_{\Delta h_i \Delta y} \tag{5.4}$$

and the the contribution score of input neuron x_j to target neuron y is computed as

$$C_{\Delta x_j \Delta y} = m_{\Delta x_j \Delta y} \times \Delta x_j \tag{5.5}$$

Shrikumar *et al.* (2017) introduces some rules including linear, rescale and reveal cancel to compute the multiplier for each neuron to its immediate inputs. These rules are suitable for activations with linear functions or nonlinear functions with only

one input. Non-linear functions with multiple inputs are not addresses in Shrikumar *et al.* (2017), and the public implementation of DeepLIFT uses the gradient for such functions (Ancona *et al.* (2017)).

DeepSHAP modifies DeepLift by computing SHAP values for smaller components of neural network analytically and propagate them backward through the network using DeepLift's multipliers. Also, DeepSHAP uses multiple background samples instead of one reference and averages the resulting attributions with respect to one reference sample at a time (Chen *et al.* (2019)). Chen *et al.* (2019) argues that averaging SHAP values over one reference sample approaches the true SHAP values for a given background distribution.

5.2 Method: Matched Neural Network

In this section, we explain how we extend neural networks to matched case-control study designs with many variables (hundreds and thousands) for the purpose of identifying informative variables in distinguishing between case and control units. Our method is suitable for matched $1 - L$ case-control study designs with $L \geq 1$. It is built upon the idea behind Matched Forest algorithm, that is, we first transform data to a supervised setting while maintaining the matching structure of data, then, we train a neural network classifier on this transformed data set to classify each stratum and measure the importance of variables. For simplicity, we first explain how neural networks are extended to matched case-control study designs with one case and one control (matched $1 - 1$ design), then, we will extend the method to a more general matched $1 - L$ study design where each stratum has one case and $L > 1$ controls.

5.2.1 *Matched* $1-1$ *Study Design*

We use the notation in Shomal Zadeh *et al.* (2020) to describe the problem. Consider a matched case-control data set D with N strata consisting of one case and one control, R exposure variables denoted by $x_1, x_2, \ldots, x_R$ and M matching variables denoted by $v_1, v_2, \ldots, v_M$. The case and control values of an exposure variable x_r for stratum i are shown by $x_r^1(i)$ and $x_r^0(i)$, respectively. We transform data to a supervised setting that accounts for the matched structure of data using the data transformation method in Shomal Zadeh *et al.* (2020). That is, new case and control variables $x_r^{*1}(i)$ and $x_r^{*0}(i)$ are created for each exposure variable x_r as

$$x_r^{*k}(i) = \begin{cases} x_r^k(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ x_r^{1-k}(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.6}$$

for $r \in \{1, 2, \ldots, R\}$. The first N rows are the original pairs and the second N rows (referred as counterfactual) are new matched pairs for each of which the exposure values of case and control are interchanged. Variable d_r^* is also created for each numerical exposure variable as

$$d_r^* = x_r^{*1} - x_r^{*0} \tag{5.7}$$

to help identify the correct effect. The method also generates new columns for matching variables as

$$v_m^+(i) = \begin{cases} v_m(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ v_m(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.8}$$

A label is also defined for each matched pair as

$$y(i) = \begin{cases} 1 & \text{for } i \in \{1, 2, \ldots, N\}, \\ 0 & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.9}$$

to distinguish between observed values and counterfactuals. If all R exposure variables in D are numerical, this transformation will create new matched case-control data set D^* with $2N$ instances and $M + 3R + 1$ columns.

We train a neural network classifier on this transformed data set D^* to classify data and measure the importance of each variable in predicting the class probabilities. Our method uses DeepSHAP to explain the predicted output and assign SHAP values to each individual variable. SHAP value of a variable is a measure of its importance score. It measures the impact of each variable for every single prediction. A positive (negative) SHAP score of a variable indicates that it increases (decreases) model's prediction. In matched case-control analysis, we are interested in the importance of each matching variable v_m and each exposure variable x_r that has three associated columns in D^*, namely, x_r^{*1}, x_r^{*0} and d_r^*. Thus, to compute an importance score for the exposure variable x_r, we need to aggregate the SHAP scores of its three associated variables. As DeepSHAP is an additive feature attribution method, a summation of these scores would represent the whole impact of exposure variable x_r on model's prediction. Let VI denote the original SHAP values computed by DeepSHAP method and let MNI be the Matched Neural Network's importance score which is a modification of VI scores for the matched dataset.

The MNI score of a matching variable v_m for stratum i is computed as

$$MI(v_m(i)) = VI(v_m^+(i)) \tag{5.10}$$

for $m \in \{1, 2, \ldots, M\}$ and $i \in \{1, 2, \ldots, 2N\}$. The MNI score of each exposure variable x_r is computed as

$$MI(x_r(i)) = VI(x_r^{*1}(i)) + VI(x_r^{*0}(i)) + VI(d_r^*(i)) \tag{5.11}$$

for $r \in \{1, 2, \ldots, R\}$ and $i \in \{1, 2, \ldots, 2N\}$. The global score for each matching and exposure variable is then computed by taking the average of absolute values of

MNI scores over the entire strata in the D^*. That is, we use $\frac{1}{2N}\sum_{i=1}^{2N}|MNI(x_r(i))|$ and $\frac{1}{2N}\sum_{i=1}^{2N}|MNI(v_m(i))|$ as the global importance score of exposure variable x_r for $r \in \{1, 2, \ldots, R\}$ and matching variable v_m for $m \in \{1, 2, \ldots, M\}$, respectively.

5.2.2 Matched $1 - L$ Study Design

Here, we extend our method to matched case-control study designs with $L > 1$. Similar to the notations used in section 5.2.1, let N be the number of strata, $\{x_1, x_2, \ldots, x_R\}$ denote R exposure variables, and $\{v_1, v_2, \ldots, v_M\}$ denote M matching variables. For stratum i, we show the case value for exposure variable x_r by $x_r^1(i)$ and its L controls by $x_r^{01}(i), x_r^{02}(i), \ldots, x_r^{0L}(i)$. Our method first creates L difference variables $\{d_r^1, d_r^2 \ldots, d_r^L\}$ for each exposure variable x_r as

$$d_r^l(i) = x_r^1(i) - x_r^{0l}(i)$$

for $r \in \{1, 2, \ldots, R\}$, $l \in \{1, 2, \ldots, L\}$ and $i \in \{1, 2, \ldots, N\}$. Generally, for an exposure variable that has an effect, its associated difference variables should be significantly larger (or smaller) than 0 for many strata. Our method transforms this problem to a supervised setting and let the classifier detect important variables with an effect. The key idea is to keep all the available information corresponding to a stratum (matching, case and control values) in one row of data set and transform data to a supervised setting which enables the classifiers to detect important variables. Our method transforms data by creating new L difference variables $\{d_r^{*1}, d_r^{*2} \ldots, d_r^{*L}\}$ for each exposure variable x_r as

$$d_r^{*l}(i) = \begin{cases} d_r^l(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ -d_r^l(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.12}$$

for $r \in \{1, 2, \ldots, R\}$ and $l \in \{1, 2, \ldots, L\}$. A label is also defined for each stratum as

$$y(i) = \begin{cases} 1 & \text{for } i \in \{1, 2, \ldots, N\}, \\ 0 & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.13}$$

to distinguish between the original differences in the first N rows and their negative values in the second N rows. If an exposure variable is important, the classifier should be able to distinguish between different labels and identify its effect. To enable the method to detect nonlinear and interaction effects between exposures, variables x_r^{*1} and x_r^{*0} are generated for each exposure variable x_r as

$$x_r^{*1}(i) = \begin{cases} x_r^1(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ \dfrac{1}{L}\displaystyle\sum_{l=1}^{L} x_r^{0l}(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.14}$$

and

$$x_r^{*0}(i) = \begin{cases} \dfrac{1}{L}\displaystyle\sum_{l=1}^{L} x_r^{0l}(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ x_r^1(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{5.15}$$

for $r \in \{1, 2, \ldots, R\}$. Equations 5.14 and 5.15 are similar to Equation 5.6, but here, we use the average of L control values of an exposure within a stratum. New matching variables $\{v_1^+, v_2^+, \ldots, v_M^+\}$ are also generated similar to Equation 5.8 to help identify interaction effects between matching and exposure variables. This data transformation method will create new data set D^* with $2 \times N$ rows and $M + (L + 2) \times R + 1$ columns.

Similar to our method for matched $1-1$ study design, we train a classifier, here neural network, on this transformed data set to classify strata and to identify important variables. The importance of each variable is measured through SHAP values which are estimated by the DeepSHAP algorithm. The matched neural network's

importance score of exposure variable x_r for stratum i is computed as

$$MNI(x_r(i)) = VI(x_r^{*1}(i)) + VI(x_r^{*0}(i)) + \sum_{l=1}^{l=L} VI(d_r^{*l}(i)) \tag{5.16}$$

for $r \in \{1, 2, \ldots, R\}$ and $i \in \{1, 2, \ldots, 2N\}$. The matched neural network importance score of matching variable v_m^+ for stratum i is computed as

$$MNI(v_m(i)) = VI(v_m^+(i)) \tag{5.17}$$

for $m \in \{1, 2, \ldots, M\}$ and $i \in \{1, 2, \ldots, 2N\}$. To measure the global importance score of a variable across all strata, we use $\frac{1}{2N}\sum_{i=1}^{2N} |MNI(x_r(i))|$ for exposure variable x_r, $r \in \{1, 2, \ldots, R\}$, and $\frac{1}{2N}\sum_{i=1}^{2N} |MNI(v_m(i))|$ for matching variable v_m, $m \in \{1, 2, \ldots, M\}$.

5.3 Experiments

Here, we demonstrate the effectiveness of our method in identifying important variables from high-dimensional matched case-control studies through simulations and a biomedical application. We also compared the performance of our method with Matched Forest and Conditional Classification method.

5.3.1 Simulation Study

We simulated matched case-control data sets to evaluate the performance of our method and compared it with Matched Forest (Shomal Zadeh *et al.* (2020)) and the conditional classification method (Stanfill *et al.* (2019)). The conditional classification method proposes a data preprocessing approach and then trains a classifier on this transformed data to classify case and control instances and identify important exposure variables. In our experiments, we used neural network algorithm for the conditional classification method and refer to it as conditional neural network. Their

method did not provide any explanation of how matching variables are handled, so we include them in the transformed data without any transformation of their values.

For each simulation study, 10 matched case-control data sets are generated with 600 strata, 5 matching and 100 exposure variables. The number of controls matched to each case (L) ranges from 1 to 4, and it is constant across strata for each simulation study. The data sets are simulated with different types of effects including linear, interaction between a matching and an exposure and interaction between two exposure variables. Data sets with interaction effects are simulated based on XOR rule such that each variable individually does not have any marginal effect, but its combination with other variables shows an effect. Our data generation process is motivated by Shomal Zadeh *et al.* (2020). Each matching variable unless otherwise stated is generated randomly from Poisson ($\lambda = 5$) distribution. To generate the case and control variables of exposure variable x_r, unless otherwise stated, the case values (x_r^1) are first generated according to the Uniform distribution between 1 and 50 ($U(1, 50)$) and values of each control variable (x_r^{0l}, for $l \in \{1, 2, \ldots, L\}$) are generated according to $x_r^{0l} = x_r^1 - d_r^l$ where d_r^l is normally distributed $N(\mu_r^l, 1)$. In our simulations, $\mu_r^l \in \{1, 1.5, 2\}$ for exposure variables with positive effect, $\mu_r^l \in \{-1, -1.5, -2\}$ for exposure variables with negative effect and $\mu_r^l = 0$ for exposure variables without any effect. Let $|\mu_r^l|$ be the effect size. Larger values of $|\mu_r^l|$ indicate a stronger effect.

The neural network architecture used in our proposed method and the conditional classification algorithm consists of 3 fully connected layers of size 30 with Relu activation followed by the softmax output layer with 2 nodes. To handle overfitting due to large number of variables, we used $L1$ regularization with $\lambda \in \{0.005, 0.007\}$ on the weights connecting input layer to the first hidden layer. The networks were trained using Adam optimization algorithm with learning rate of 0.01. The number of epochs is set to 120 and batch size to 5. We used 5 fold cross validation to split the strata

into training and test sets, and applied DeepSHAP with 600 background samples to estimate the effect of each variable for instances in each test set. DeepSHAP method provides SHAP scores corresponding to each output node of the model. When the label is binary, the SHAP values of the 2 output nodes have the same value but different sign. In our simulation results, we report only SHAP values estimated for the first output node that takes 1 for the observed strata and 0 for the counterfactuals.

To train Matched Forest, we set number of trees to 1000, number of variables selected at each split to $\sqrt{(p)}$ where p is total number of variables in the transformed data and grow trees to purity. We also used 5 fold cross validation to split strata into training and test sets. Matched Forest was trained on the training portion and the average of Matched Forest importance scores (MFI) over 10 folds was used for the importance of each variable. MFI scores are based on the Gini importance measure of Random Forest algorithm.

Linear Effect

Here, we simulated matched case-control data sets with $L = 2$ such that exposure variable x_1 has a positive linear effect and the remaining exposure variables $x_r \in \{x_2, x_3, \ldots, x_R\}$ have no effect. That is, we set $\mu_1^l = 1$ for $l \in \{1, 2\}$ and $\mu_r^l = 0$ for $r \in \{1, 2, \ldots, R\}$ and $l \in \{1, 2\}$.

Figures 5.1a, 5.1b and 5.1c show respectively MNI scores from Matched Neural Network, MFI scores from Matched Forest and SHAP values from Conditional Neural Network for exposure variables. For brevity, we only show the top 20 most important exposure variables from each method. We observe that all three methods detect the effect of x_1 because this variable has received the highest importance value by all the methods. Figure 5.1d shows a summary of SHAP values obtained from one replicate of matched neural network for the top 5 most important variables. Each

point in this plot is the estimated SHAP value for each instance and each variable in the transformed data set D^*. The y axis represents the variables sorted based on their global importance (the average of SHAP value magnitudes over all samples) and the x axis shows the SHAP values. The color represents the values of features from the smallest (blue) to the largest (red). From this figure, we can observe that the difference variables d_1^{*1} and d_1^{*2} associated with exposure variable x_1 are the most important features in distinguishing between observed strata and counterfactuals, and their SHAP values also increase with their feature values in D^*. We also tested the change in the accuracy of matched neural network by changing the number of controls matched to each case (L) and the effect size ($|\mu_1^l|$ for $l \in \{1, 2, \ldots, L\}$). The results are shown in Figure 5.1e for $L \in \{1, 2, 3, 4\}$ and $|\mu_1^l| \in \{1, 1.5, 2\}$. There can be seen that the accuracy of matched neural network increases with larger L and effect size.

Interaction Effect Between A Matching And An Exposure

Here, data sets are simulated with an interaction effect between matching variable v_1 and exposure variable x_1 and no effect for other exposure variables. The number of controls matched to each case is set to 2. Exposure variable x_1 is generated with a positive effect when $v_1 \leq 5$ and a negative effect when $v_1 > 5$. Specifically, for $N/2$ of the strata, $\mu_1^l > 0$ for $l \in \{1, 2\}$ and v_1 is generated from $U(1, 5)$ and for the remaining $N/2$ strata, $\mu_1^l < 0$ for $l \in \{1, 2\}$ and v_1 is generated from $U(5, 10)$.

Figure 5.2 shows variable importance measures of matching and exposure (top 20) variables from matched neural network (Figures 5.2a and 5.2b), matched forest (Figures 5.2c and 5.2d), and the conditional neural network model (Figures 5.2e and 5.2f). Both matched neural network and conditional neural network models detect the interaction effect between variables v_1 and x_1 because these two variables have received significantly larger importance scores than other matching and exposure

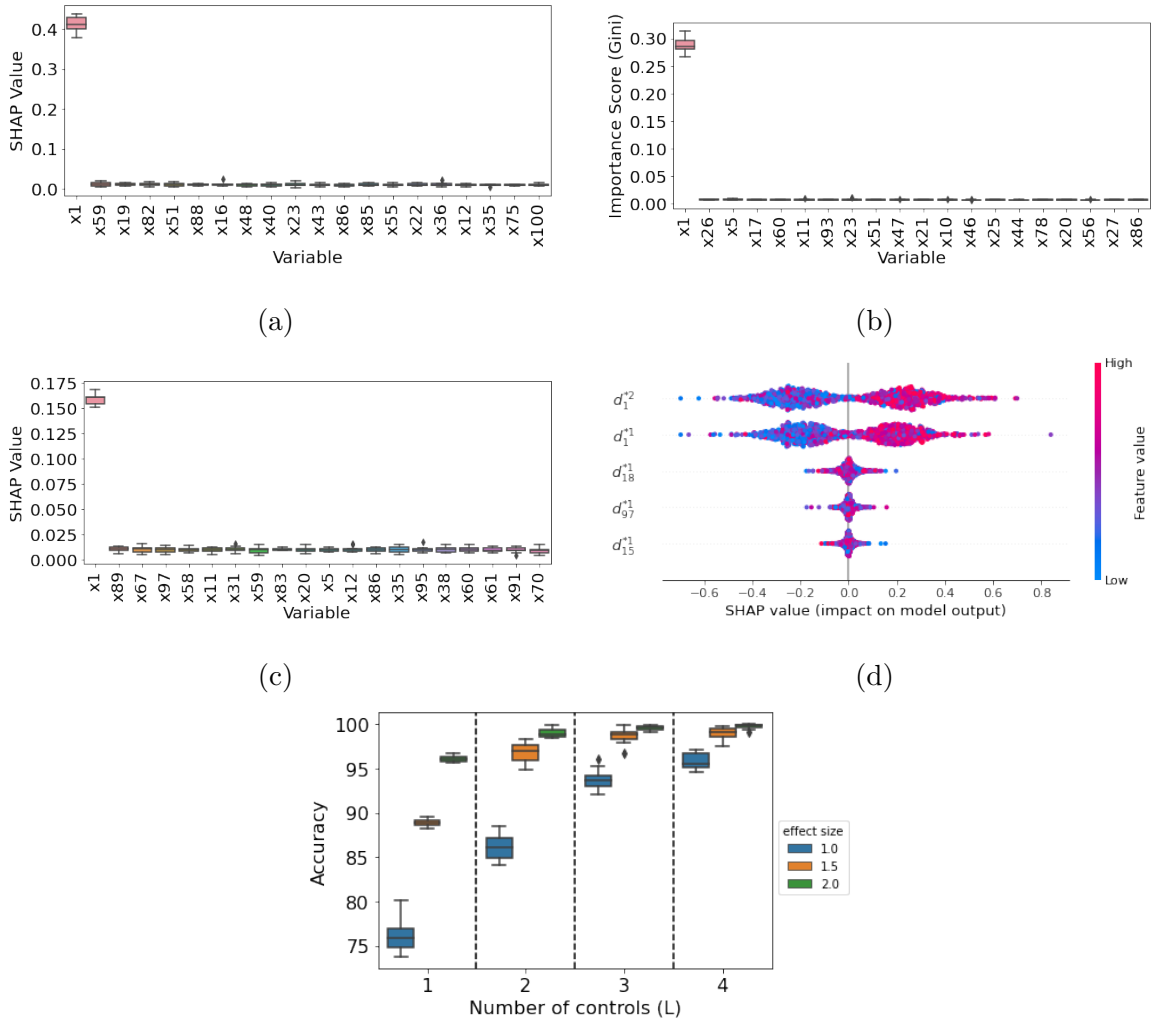


Figure 5.1: Linear Effect: (a) MNI Scores (Shap Values) of Exposure Variables from Matched Neural Network. (b) MFI Sores of Exposure Variables from Matched Forest. (c) SHAP Values of Exposure Variables from Conditional Neural Network. (d) Summary Plot of SHAP Values Estimated by Matched Neural Network. Each Point Corresponds to a Variable and an Instance, and the Color Represents the Feature Values from Low (Blue) to High (Red). (e) Effect of l and Effect Size ($|\mu_1^l|$) on the Accuracy of Matched Neural Network.

variables. However, matched forest only detects the effect of matching variable v_1 and it fails to detect the effect of x_1 because it is not in the list of top 20 most important variables shown in Figure 5.2c.

Figure 5.3a shows how the impact of variable d_1^{*1} on model's output changes as its value varies across strata. Each point in this plot corresponds to a stratum. The x-axis shows the feature value of d_1^{*1} and y-axis shows its corresponding SHAP value. The color represents the feature value of v_1 in D^* from the smallest value (blue) to largest (red). This plot also shows how these two variables are interacted. For example, for strata with $d_1^{*1} > 0$, the impact of d_1^{*1} on model's output is larger when $V1 < 5$, and it increases with larger values of d_1^{*1}. We also tested the impact of number of controls (L) and effect size ($|\mu_1^l|$ for $l \in \{1, 2, \ldots, L\}$) on the accuracy of matched neural network. The results are shown in Figure 5.3b for $L \in \{1, 2, 3, 4\}$ and $|\mu_1^l| \in \{1, 1.5, 2\}$. The accuracy of matched neural network increases as $|\mu_1^l|$ and L become larger.

Interaction Effect Between Two Exposure Variables

In this section, we explain our simulation study for matched data sets with an interaction effect between exposure variables x_1 and x_2, and no effect for other exposure variables $\{x_3, x_4, \ldots, x_R\}$. The number of controls matched to each case is set to 2. Data sets are generated such that x_1 has a positive effect when $x_2^{*1} < 25$ and a negative effect when $x_2^{*1} \geq 25$. That is, for $N/2$ strata, $\mu_1^l = 1$ for $l \in \{1, 2\}$ and x_2^{*1} is generated from $U(1, 25)$, and for the remaining strata, $\mu_1^l = -1$ for $l \in \{1, 2\}$ and x_2^{*1} is generated from $U(25, 50)$. For exposure variables $\{x_2, x_3, \ldots, x_r\}$, $d_r^{*l} = 0$ for $l \in \{1, 2\}$.

Figures 5.4a, 5.4b and 5.4c show variable importance scores of the top 20 exposure variables from matched neural network, matched forest and conditional neural net-

work, respectively. Matched neural network detects the simulated interaction effect between x_1 and x_2, because they both have received significantly larger scores than other variables. For matched forest, although x_1 and x_2 are in the top 3 variables, they do not have significantly larger scores than other exposure variables without any effect, thus, matched forest only weakly detects this interaction effect. Conditional neural network also fails to detect the interaction effect because neither x_1 nor x_2 are seen in the top 20 variables shown in Figure 5.4c.

Figure 5.4d shows how the impact of d_1^{*1} on model's output changes with its feature value and the value of x_2^{*1}. When $d_1^{*1} > 0$, its impact on models output is larger for $x_2^{*1} < 25$ than $x_2^{*1} > 25$. Also, when $x_2^{*1} < 25$, the impact of d_1^{*1} increases as its feature value becomes larger, and when $x_2^{*1} > 25$, the impact of d_1^{*1} increases as its feature value becomes smaller. We also tested the effect of number of controls (L) and effect size of exposure variable x_1 ($|\mu_1^l|$ for $l \in \{1, 2, \ldots, L\}$) on the accuracy of matched neural network. Figure 5.4e shows the results for $L \in \{1, 2, 3, 4\}$ and $|\mu_1^l| \in \{1, 1.5, 2\}$. We can see in this Figure that the accuracy of matched neural network generally improves when L and $|\mu_1^l|$ increase. But there can be seen that the accuracy of MNN decreases when number of controls changes from 3 to 4 for the effect size is of 2. Thus, number of controls matched to a case should be selected with cautious.

Summary of Simulation Results

We observed in our simulations that MNN performs better than MF and Conditional Neural Network in identifying variables with interaction effects. However, they all perform comparably when important variables have linear effects. Also, generally, when number of controls matched to a case unit increases, the accuracy of MNN improves. But the accuracy might also decrease when number of controls is very

large. Thus, the number of controls matched to each case should be selected with cautious.

5.4 Case Study: Childhood Acute Lymphoblastic Leukemia Study

In this section, we utilize Childhood Acute Lymphoblastic Leukemia data set (Bhojwani *et al.* (2006)) to show the effectiveness of matched neural network in identifying important variables. This data set includes 35 matched pairs where each represent a child with acute lymphoblastic leukemia who were relapsed after therapy. Each pair consists of $22,283$ gene expression profiles in bone marrow measured at the diagnosis (Case) and relapsed stage (control). The objective of this study is to identify gene expression profiles significantly different between cases and controls.

We trained matched neural network on this data set to evaluate the importance of each gene expression profile, and compared its performance with Matched Forest. The neural network architecture consists of 3 hidden layers of size 50 with Relu activation function followed by the softmax output layer with 2 nodes. To reduce the risk of overfitting due to the large number of exposure variables $(22,283)$, $L1$ regularization with parameter $\lambda = 0.005$ is used for the weights connecting the input layer to the first hidden layer. We used stochastic gradient descent as the optimization algorithm for training neural network and set learning rate to 0.001, weight decay parameter to 0.01, and momentum to 0.9. The number of epochs is set to 30 and the batch size to 5. We also used 10 fold cross validation to separate data into training and test, and to account for the randomness in cross validation, this training procedure was repeated 10 times. Matched Forest was also conducted in 10 replicates with $10,000$ trees and 10 fold cross validation. The number of variables selected at each split is also set to $(\sqrt{(22283)})$ and trees were grown to purity.

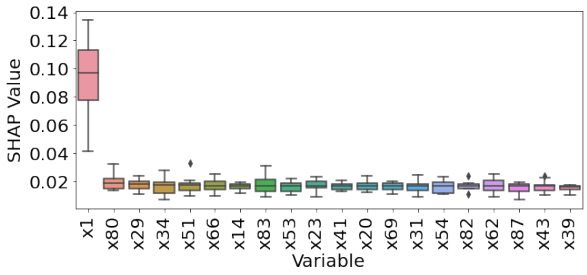
Figure 5.5a shows the median of MNI scores over the 10 replicates of matched neural network for the top 50 genes with largest importance scores. These variables were ranked from the most (rank 1) to the least (rank 22, 283) importance based on their importance values. There are a number of differences between variables ranking from matched neural network and Matched Forest. Among the top 100 most important variables selected by Matched Neural Network, 41 of them are also ranked between 1 to 100 by Matched Forest. Some of the most important genes selected by both methods are 204642_at, 205167_s_at, 200980_s_at and 217491_x_at which are ranked $3rd$, $2nd$, $4th$ and $7th$ by Matched Neural Network respectively and ranked $1st$, $9th$, $5th$ and $33rd$ by Matched Forest, respectively. There are also some genes which are ranked high by Matched Neural Network, but received lower ranks by Matched Forest. For example, 202867_s_at is ranked $18th$ by Matched Neural Network and $234th$ by Matched Forest. The median of accuracy over 10 runs of Matched Neural network is 78 which is lower than the median of accuracy for Matched Forest which is 83. We think that lower accuracy of Matched Neural Network is due to small number of matched pairs (only 35), because Neural Network models usually work better when a large amount of data is available.

As MNI scores are computed for each instance individually, we can plot MNI scores of a variable across all matched pairs and see how its impact changes as its value varies. For example, we can see in Figure 5.5b for exposure variable 208511_at (ranked $17th$) that the impact of its associated difference variable $d^*_{208511\_at}$ increases as its value gets larger. However, we can see a decreasing trend in Figure 5.5c for exposure variable 217099_s_at (ranked $15th$) which indicates that the impact of its difference variable $d^*_{217099\_s\_at}$ decreases as its value becomes larger.

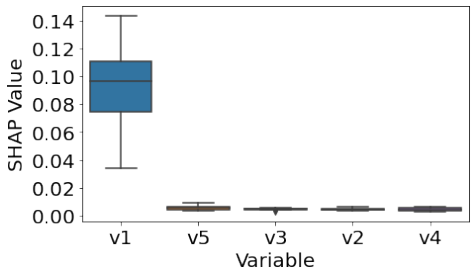
5.5 Conclusion

We proposed Matched Neural Network that identifies important matching and exposure variables from high-dimensional matched case-control data sets with hundreds and thousands of variables. This method is suitable for matched case-control study designs where each case is matched to a fixed number of controls. Matched case-control data sets are first transformed to a supervised setting while accounting for its matched structure and then a Neural Network classifier is trained on this data to identify important variables. We used a modification of SHAP values to compute the importance of each matching and exposure variable.

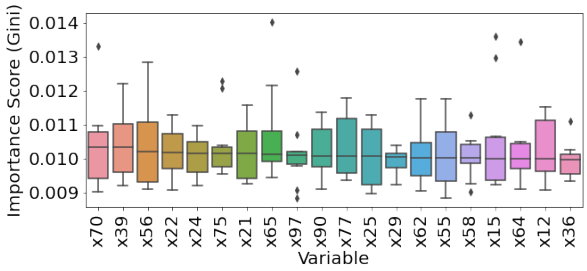
We compared the performance of Matched Neural Network with alternative variable selection methods including Matched Forest and Conditional Classification through simulation studies and a biomedical application. We observed in our simulations that Matched Neural Network performs better than alternative methods to identify interaction effects. However, when number of strata is small, we observed in the analysis of biomedical data set that Matched Neural Network does not perform as well as Matched Forest. Also, SHAP values enable us to see how the impact of a variable varies depending on its values and the value of an interacting variable.



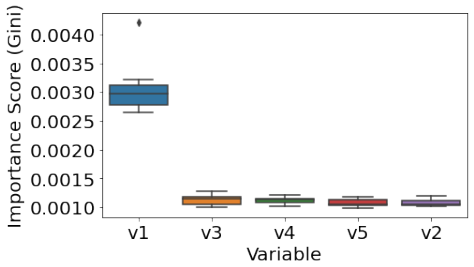
(a) Matched neural network: MNI scores of exposure variables



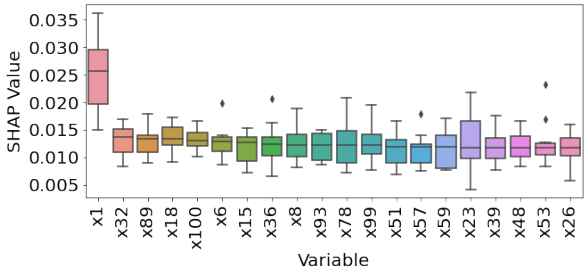
(b) Matched neural network: MNI scores of matching variable



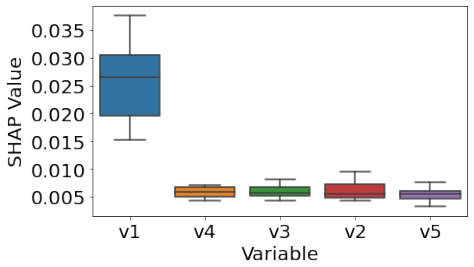
(c) Matched forest: MFI scores of exposure variables



(d) Matched forest: MFI scores of matching variables

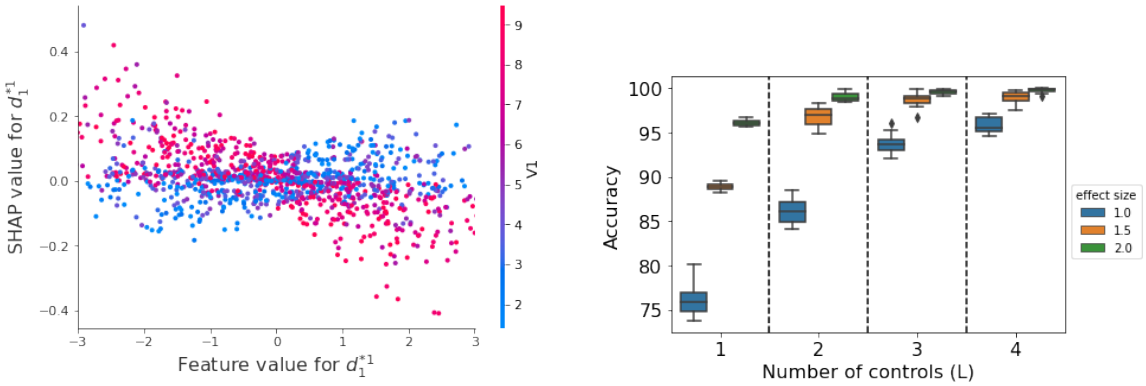


(e) Conditional neural network: SHAP values of exposure variables



(f) Conditional neural network: SHAP values of matching variables

Figure 5.2: Interaction between Matching Variable v_1 and Exposure Variable x_1: (a) MNI Scores (SHAP Values) of Exposure Variables from Matched Neural Network. (b) MNI Scores (SHAP Values) of Matching Variables from Matched Neural Network.(c) MFI Sores of Exposure Variables from Matched Forest. (d) MFI Scores of Matching Variables from Matched Forest.e SHAP Values of Exposure Variables from Conditional Neural Network. (f) SHAP Values of Matching Variables from Conditional Neural Network.



(a) (b)

Figure 5.3: Interaction between Matching Variable v_1 and Exposure Variable x_1: (a) SHAP Dependence Plot to Show the Effect of d_1^{*1} Across All Data and Its Interaction with Matching Variable v_1 (b) Effect of l and Effect Size ($|\mu_1^l|$) on the Accuracy of Matched Neural Network.

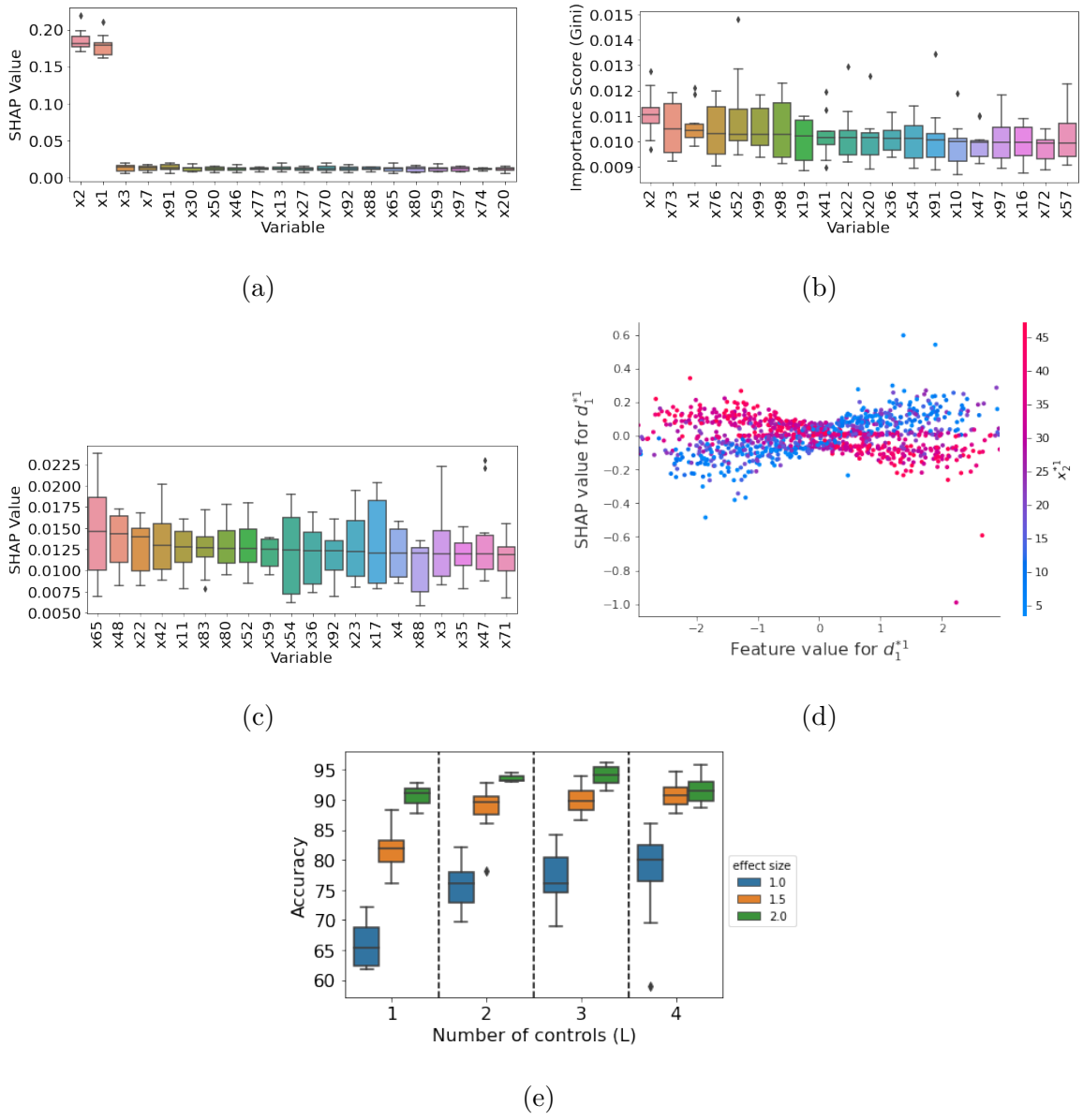
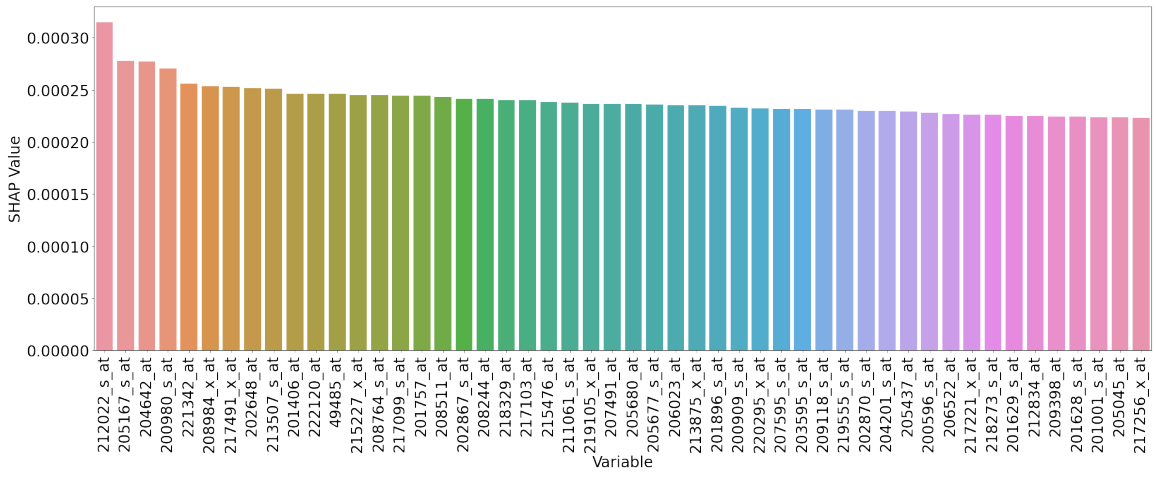
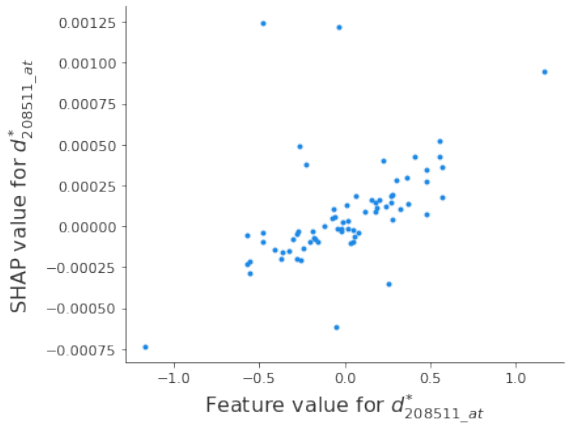


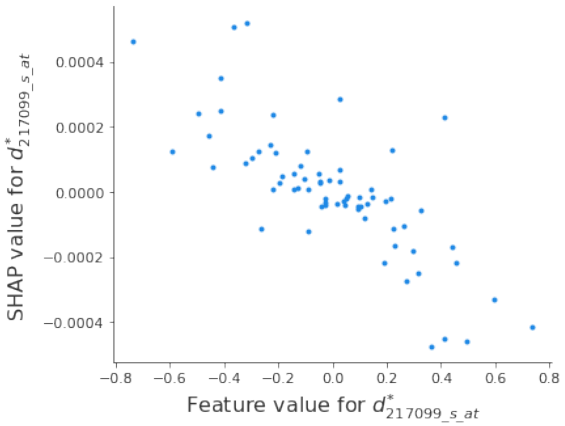
Figure 5.4: Interaction between Exposure Variables x_1 and x_2: (a) MNI Scores (SHAP Values) of Exposure Variables from Matched Neural Network. (b) MFI Scores of Exposure Variables from Matched Forest. (c) SHAP Values of Exposure Variables from Conditional Neural Network. (d) Summary Plot of Shap Values Estimated by Matched Neural Network. Each Point Corresponds to a Variable and an Instance, and the Color Represents the Feature Values from Low (Blue) to High (Red). (e) Effect of L and Effect Size ($|\mu_1^l|$) on the Accuracy of Matched Neural Network.



(a)



(b)



(c)

Figure 5.5: Childhood Acute Lymphoblastic Leukemia Study: (a) The Median of MNI Scores (SHAP Values) over the 10 Replicates of Matched Neural Network for the Top 50 Genes with Largest Importance Scores. (b) The Impact of Variable $d^*_{208511\_at}$ Versus Its Value Across All Data. (c) The Impact of Variable 217099_*s_at* Versus Its Value Across All Data.

Chapter 6

MULTINOMIAL MATCHED LEARNER: SUPERVISED MACHINE LEARNING FOR HIGH-DIMENSIONAL MATCHED STUDIES WITH MULTIPLE LEVELS OF THE OUTCOME

6.1 Introduction

In many applications, observations within a data set have some dependency in their structure. They usually form groups of observations where observations in a group are correlated. Examples include longitudinal data sets where repeated observations taken from a unit at different time points or occasions, variables measured for a same subject before and after of an intervention (e.g., treatment), and matched study designs where observations are grouped together based on some baseline variables. The analysis of this kind of data sets requires techniques to handle dependency among observations to achieve valid inference.

In this research, we consider a matched study design with the objective of identifying important variables associated with a nominal outcome with more than two levels. In this study design, each stratum consists of three or more units where each is from one of the outcome levels. Units per stratum differ with respect to their outcome level, but similar with respect to certain baseline variables used for matching (for example, age and gender). These baselines variables used to create matched samples are also referred as matching variables. For each unit within a stratum, a large number (hundreds or thousands) of exposure variables are measured. For example, in a clinical application, these variables may include patient characteristics, health

conditions, or genes. The objective of this study is to assess the relationship between exposure variables and the nominal outcome of interest.

Matching is commonly performed in observational studies to increase efficiency by equating the distribution of baseline variables in different outcome levels (Rothman *et al.* (2008)). Matching is performed on variables believed to be confounders; variables which are associated with both exposure variables and the outcome of interest (Rose and Van der Laan (2009)). If the matching variable is only associated with the outcome, there is a loss of efficiency compared to an unmatched study design. Also, if the matching variable is only associated with the exposure, but not the outcome, the variance of estimator will increase. We should also avoid selecting variables for matching which are along the causal pathway between exposure variables and the outcome of interest, because this will create bias in our estimation (Stuart (2010) and Rose and Van der Laan (2009)). Variables which are typically used for matching in epidemiological studies are demographic variables such as age, sex, race, etc. An ideal method for matching is the exact matching method where units within each matched set have the equal values for all matching variables (Stuart (2010)). Another commonly used matching method is the nearest neighbor matching where units within a stratum have the smallest distance from each other (Stuart (2010)). In this research, we assume that matched sets or strata are created by the exact matching method. The number of units from each outcome level within each stratum can vary in general. In this research, we assume that each stratum has one unit from each outcome level.

Traditional methods for analysing this type of study design are not suitable for high-dimensional matched data sets with hundreds or thousands of variables and complex relationship among them. Majority of these methods are based on the conditional logistic regression (CLR) (Hosmer Jr *et al.* (2013)) model which uses a linear

logistic regression model and a conditional likelihood approach to handle the matched structure of data set. To assess interaction effects among variables, interaction terms (products of two or more variables) are included in the CLR model. However, this increases the dimensionality of data set and makes this model intractable.

These limitations of the CLR based models have motivated us to use a machine learning model for matched study designs; a model which inherently handles high-dimensionality, complex nonlinear and interaction effects. We present a new machine learning algorithm, Multinomial Matched Learner (MML) to identify important variables from high-dimensional matched data with multiple levels for the outcome. Our method is motivated by the Matched Forest (MF) algorithm developed by Shomal Zadeh *et al.* (2020) for matched case-control study designs where outcome has two levels. The main idea of our method is to transform data set to a supervised setting based on the potential outcome model (Neyman (1923), Rubin (1977)) while accounting for the matched structure of data set. Similar to MF, for each unit within a stratum, we estimate its counterfactual which is defined for the case (control) unit as its potential exposure value if the unit was a control (case). Then a label is defined for each stratum and a classifier with modified variable importance score is used to identify important variables. One advantage of our method is that any classifier can be applied. We used Random Forest and Neural Network in our experiments because both handle high-dimensionality, nonlinear and interaction effects in high-dimensional setting.

Section 6.2 presents previous work on matched study designs. Section 6.3 describes our method. Section 6.4 presents and discusses simulation and case studies. Finally, section 6.5 concludes this work.

6.2 Background

6.2.1 Related Work

Th predominant method in the literature for analyzing matched study designs with binary outcome is Conditional Logistic Regression (CLR) and its variants for high-dimensional setting for example Balasubramanian *et al.* (2014), Asafu-Adjei *et al.* (2017), and Qian *et al.* (2014). These models all use a linear model which is supplemented by interaction terms to assess interactions among variables. Thus, they are not suitable for matched data sets with hundreds or thousands of variables.

There are also other references which are not based on CLR. For example, Adewale *et al.* (2010) proposed two versions of boosting algorithms for data sets with correlated binary outcome levels. The first algorithm uses gradient boosting algorithm with weighted least square loss function that handles correlation among outcomes. In the second algorithm a likelihood optimization boosting algorithm is modified by using a generalized linear mixed model. Both algorithms proposed by Adewale *et al.* (2010) use a linear model which requires interaction terms to assess interactions among variables. Thus, they have difficulty identifying non-linear and interaction terms in high-dimensional matched data sets.

Another research in this domain is by Stanfill *et al.* (2019) who proposed a data transformation method which transforms exposure variables to their null space and applies any classification method on this transformed data set to identify important exposure variables. This method does not handle dependency among units within a stratum, which is recommended by statistical principles. It breaks each stratum into multiple instances which are known to be dependent. We observed in our experiments in Chapter 5 that this method has difficulty identifying interaction effects between exposure variables.

Matched Forest (MF) proposed by Shomal Zadeh *et al.* (2020) is also a recent method for identifying important variables from matched data sets with binary outcome. This method uses a data transformation method based on the potential outcome model (Splawa-Neyman *et al.* (1990) and Rubin (1974)) to convert matched data sets to supervised setting while accounting for their matched structure. Then, it applies a classifier on the transformed data set to identify important matching and exposure variables. This method does not have the limitations of methods described previously. It inherently handles high-dimensionality, nonlinear and interaction effects. Section 6.2.2 explains Matched Forest in more details because this is the basis of our method.

Research on matched data sets with more than two outcome levels is limited. The traditional method is to use the binary CLR method for each pair of outcomes (Liang and Stewart (1987)). Also, there are some references that use a joint modeling of all outcome levels to estimate the coefficients of exposure variables. For example, Liang and Stewart (1987), Becher and Jöckel (1990) and Gebregziabher *et al.* (2010) extend CLR to matched case-control data sets where either case or control units are selected from multiple groups. Each group is considered as an outcome level. Mukherjee *et al.* (2007) demonstrated that the joint modeling of multiple outcomes is more efficient than using separate CLR models for each pair of outcome levels. These models also have similar limitations as CLR model for binary outcome, which makes them not suitable for high-dimensional matched data sets with hundreds or thousands of variables and complex nonlinear and interaction effects.

6.2.2 Matched Forest

Here, we provide a summary of Matched Forest (MF) algorithm which was proposed by Shomal Zadeh *et al.* (2020) for matched study designs with two possible

outcomes in each stratum. Our method is the extension of MF for matched study designs with more than two outcome levels in each stratum.

The MF algorithm first converts the matched data set into a supervised setting based on the potential outcome frame work (Splawa-Neyman *et al.* (1990) and Rubin (1974)). Second, a classifier is trained on this transformed data set to identify important variables.

Consider a matched study design with N strata and an outcome with $T = 2$ levels with no specific ordering, where $t \in \{0, 1\}$ shows the outcome level. Shomal Zadeh *et al.* (2020) considered a matched case-control study design for their analysis where each stratum has one case unit ($t = 1$) and one control unit ($t = 0$). Let $\{x_1, x_2, \ldots, x_R\}$ denote R exposure variables and $\{v_1, v_2, \ldots, v_M\}$ be M matching variables to create N strata. The value of the exposure variable x_r for the case and control units in stratum i are represented by $x_r^1(i)$ and $x_r^0(i)$ respectively for $i \in \{1, 2, \ldots, N\}$ and $r \in \{1, 2, \ldots, R\}$. It is assumed that an exact matching method is used to create each strata. The value of matching variable v_m for units in the stratum i is represented by $v_m(i)$ for $m \in \{1, 2, \ldots, M\}$ and $i \in \{1, 2, \ldots, N\}$.

MF creates new variables x_r^{*1} and x_r^{*0} for each exposure variable x_r as

$$x_r^{*t}(i) = \begin{cases} x_r^t(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ x_r^{1-t}(i - N) & \text{for } i \in \{N + 1, N + 2, \ldots, 2N\} \end{cases} \tag{6.1}$$

for $r \in \{1, 2, \ldots, R\}$ and $t \in \{0, 1\}$. That is, the first N rows for x_r^{*1} and x_r^{*0} correspond to original values of exposure for each stratum and the second N rows (referred as counterfactual) are generated by swapping the exposure values of case and control within each stratum. To help identify important variables, MF also creates a difference variable d_r^* corresponding to each exposure variable x_r as

$$d_r^* = x_r^{*1} - x_r^{*0} \tag{6.2}$$

for $r \in \{1, 2, \ldots, R\}$. Each matching variable v_m is also transformed according to Equation 6.3 for $m \in \{1, 2, \ldots, M\}$ to help identify interaction effects between matching and exposure variables.

$$v_m^+(i) = \begin{cases} v_m(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ v_m(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.3}$$

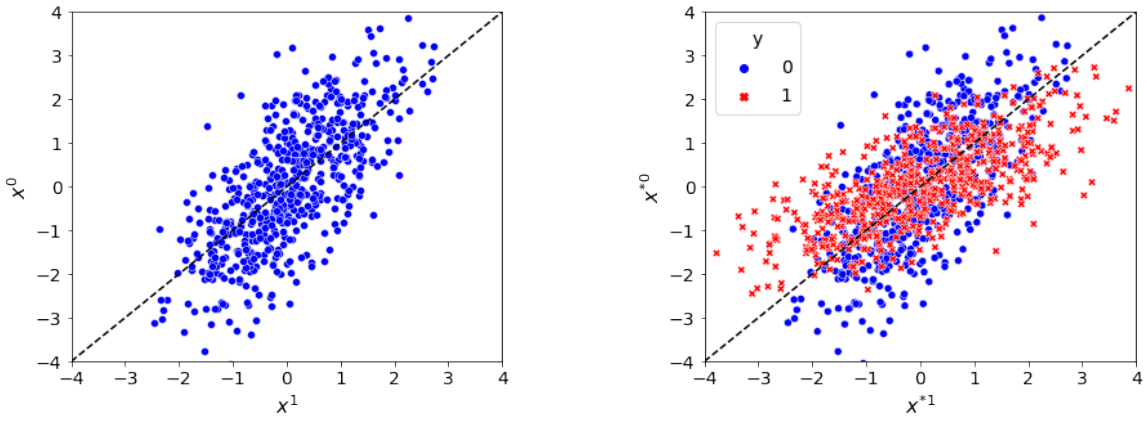
A key step is to associate a label with each pair and its counterfactual as

$$y(i) = \begin{cases} 0 & \text{for } i \in \{1, 2, \ldots, N\}, \\ 1 & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.4}$$

to distinguish between original strata and counterfactuals. A random forest classifier with a modified importance score for matched data sets is trained on this transformed data set to identify important matching and exposure variables.

Figures 6.1 and 6.2 illustrate why MF works. Figure 6.1 corresponds to a matched data set with one exposure variable x simulated with no effect. Figures 6.1a and 6.1b show respectively the original and transformed data sets for this exposure variable. Each point on these plots corresponds to a stratum with exposure values for the case and control outcomes on the x and y axes, respectively. In Figure 6.1b, the original strata are shown in blue (circle), and counterfactuals are shown in red ($\times$). In Figure 6.1a, there is no region where the value of x^1 is significantly larger or smaller than x^0 for majority of strata. In the transformed data set shown in Figure 6.1b, the original and counterfactual pairs cannot be separated and a classifier will correctly not detect any effect.

Figure 6.2a and 6.2b show respectively the original and transformed data set by MF for a simulated matched data set with one exposure variable x that has a linear effect. Each point on these plots corresponds to a stratum with exposure values for the case and control outcomes on the x and y axes, respectively. In Figure 6.2b, the



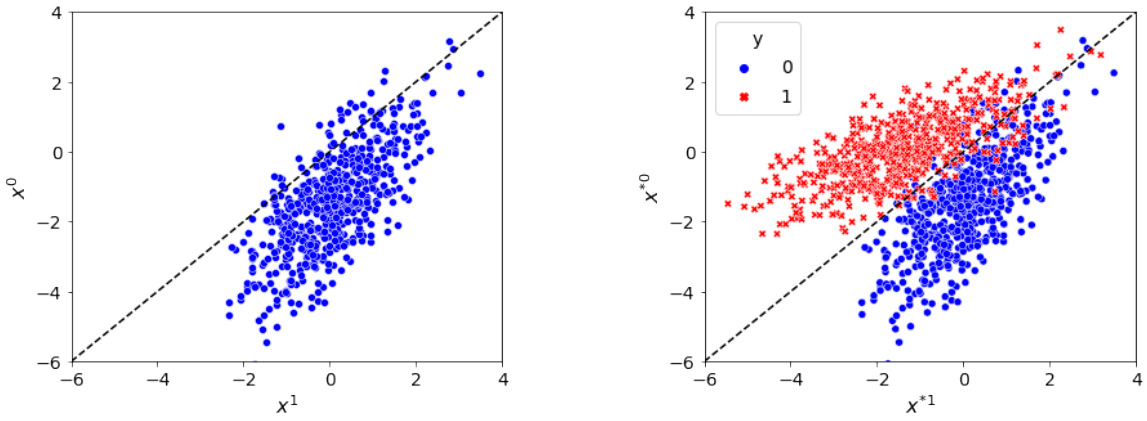
(a) (b)

Figure 6.1: The Scatter Plot of Control Versus Case for Exposure Variable x with No Effect in Original (a) and Transformed Data Set (b).

original and counterfactual pairs are shown in blue (circle) and red ($\times$) respectively. In Figure 6.2a, we can see that for majority of strata $x^1 > x^0$, hence, there is a difference between exposure values of the case and control units. In the transformed data set shown in Figure 6.2b, the original and counterfactual pairs are distinguishable. The classifier trained on this transformed data set can accurately assign labels to most of these pairs and detect the effect of the exposure variable.

6.3 Multinomial Matched Learner

We propose a new machine learning algorithm, Multinomial Matched Learner (MML), to identify important variables from high-dimensional matched data sets where outcome has more than two levels. We assume that each stratum has only one unit from each level of the outcome. Our method generalizes Matched Forest (MF) proposed in Chapter 3 to matched data sets with more than two levels for the outcome. When outcome is binary, MML is equivalent to MF. Similar to MF, our method has two major steps: (i) transform matched data set to a supervised



(a) (b)

Figure 6.2: The Scatter Plot of Control Versus Case for Exposure Variable x with Linear Effect in Original (a) and Transformed Data Set (b).

setting that accounts for the matched structure of data. (ii) train a classifier with modified importance score on this transformed data set which is able to inherently identify complex relationships (nonlinear or interaction effects) between large number of exposure and matching variables.

6.3.1 Data Transformation To Supervised Setting

Consider that the outcome of interest has more than two levels ($T > 2$) where $t \in \{0, 1, \ldots, T-1\}$ denotes the level of the outcome. Let N denote the number of strata, $\{x_1, x_2, \ldots, x_R\}$ denote R exposure variables and $\{v_1, v_2, \ldots, v_M\}$ be M matching variables. The feature matrix associated with exposure variable x_r has N rows and T columns. We denote the T columns associated with exposure variable x_r by $\{x_r^0, x_r^1, \ldots, x_r^{(T-1)}\}$ and the $T-$dimensional row vector for observed exposure values corresponding to stratum i by $x_r(i) = (x_r^0(i), x_r^1(i), \ldots, x_r^{(T-1)}(i))$ for $i \in \{1, 2, \ldots, N\}$.

Our data transformation method creates counterfactual for each stratum. Let $\tilde{x_r}(i)$ denote the counterfactual corresponding to stratum i for exposure variable x_r.

Also, let u denote the $T-$dimensional column vector with all elements equal to 1. We estimate $\tilde{x_r}(i)$ as the reflection of observed exposure for stratum i ($x_r(i)$) along the vector u. That is, $\tilde{x_r}(i)$ is obtained as

$$\tilde{x_r}(i) = x_r(i)Z \tag{6.5}$$

Where Z is the $T \times T$ reflection matrix (Z) corresponding to vector u. The Reflection matrix Z is obtained as

$$Z = 2P - I \tag{6.6}$$

where $P = u(u^T u)^{-1}u^T$ is the $T \times T$ orthogonal projection matrix onto u and I is the $T \times T$ identity matrix.

We denote the T columns associated with exposure variable x_r in the transformed data set by $\{x_r^{*0}, x_r^{*1}, \ldots, x_r^{*(T-1)}\}$. The first N rows of variables $\{x_r^{*0}, x_r^{*1}, \ldots, x_r^{*(T-1)}\}$ match with observed feature matrix corresponding to x_r and the second N rows are counterfactuals defined in Equation 6.5. Let $x_r^*(i) = (x_r^{*0}(i), x_r^{*1}(i), \ldots, x_r^{*(T-1)}(i))$ denote the $T-$dimensional feature vector associated with exposure variable x_r in the ith row of transformed data set. Thus, the vector $x_r^*(i)$ is obtained by

$$x_r^*(i) = \begin{cases} x_r(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ \tilde{x_r}(i) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.7}$$

for $r \in \{1, 2, \ldots, R\}$.

We create a new label column denoted by y in our transformed data set to distinguish between original strata and counterfactuals. The label y is defined as

$$y(i) = \begin{cases} 0 & \text{for } i \in \{1, 2, \ldots, N\}, \\ 1 & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.8}$$

If an exposure variable x_r is important, we expect the difference between its original values and counterfactuals to have large magnitude for majority of strata. Thus,

to help identify important exposure variables, for each exposure variable x_r, we create T difference variables denoted by $\{d_r^{*0}, d_r^{*1}, \ldots, d_r^{*(T-1)}\}$. The ith row of difference variables associated with exposure variable x_r is defined by the $T-$dimensional vector $d_r^*(i) = (d_r^{*0}(i), d_r^{*1}(i), \ldots, d_r^{*(T-1)}(i))$ which is defined as

$$d_r^*(i) = \begin{cases} x_r^*(i) - x_r^*(i+N) & \text{for } i \in \{1, 2, \ldots, N\}, \\ x_r^*(i) - x_r^*(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.9}$$

for $r \in \{1, 2, \ldots, R\}$.

Each matching variable v_m is also transformed to v_m^+ for $m \in \{1, 2, \ldots, M\}$ according to Equation 6.10 to enable the method to identify interactions between matching and exposure variables.

$$v_m^+(i) = \begin{cases} v_m(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ v_m(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.10}$$

Assuming that all exposure variables are numerical, this data transformation method creates a new data set with $M + 2RT + 1$ columns.

In the second step of our method, a supervised learner is trained on this transformed data set to distinguish between original strata and counterfactuals and evaluate the effect of exposure and matching variables. If an exposure variable has an effect, we would expect the supervised learner to separate the original strata and counterfactuals and identify the effect of exposure variable.

Our method compares the exposure values under T outcome levels with their average and uses a supervised learner to identify if there are any large differences between them. For an exposure variable x_r and a stratum i, the counterfactual corresponding to the observed stratum $x_r(i) = (x_r^0(i), x_r^1(i), \ldots, x_r^{(T-1)}(i))$ is $\tilde{x_r}(i) = x_r(i)Z = 2\bar{x_r}(i) - x_r(i)$ where $\bar{x_r}(i)$ is a $T-$dimensional row vector with all elements equal to $\frac{1}{T}\sum_{t=0}^{T-1} x_r^t(i)$. The difference vector $d_r^*(i)$ computed by Equation 6.9 can be

simplified and re-written as

$$d_r^*(i) = \begin{cases} 2[x_r(i) - \bar{x_r}(i)] & \text{for } i \in \{1, 2, \ldots, N\}, \\ 2[\bar{x_r}(i-N) - x_r(i-N)] & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.11}$$

for $r \in \{1, 2, \ldots, R\}$. It can be seen in Equation 6.11 that original exposure values are compared with their average over the T outcome levels.

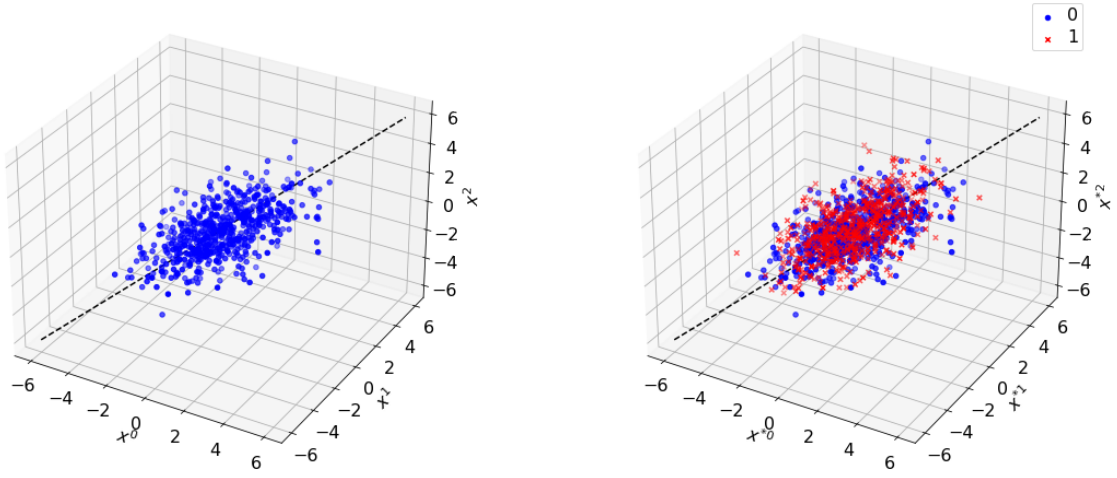
Our method is the generalization of MF for matched data sets whose outcome has more than two levels. For a binary outcome matched data set, the original exposure values corresponding to exposure variable x_r and stratum i are represented by the vector $x_r(i) = (x_r^0(i), x_r^1(i))$. MF creates counterfactual by interchanging the exposure values of the two outcome levels within each stratum. That is, the counterfactual of the original point $(x_r^0(i), x_r^1(i))$ is estimated as $(x_r^1(i), x_r^0(i))$. Our generalized method for $T \geq 2$ creates the counterfactual point corresponding to $x_r(i)$ as $\tilde{x_r}(i) = x_r(i)Z = 2\bar{x_r}(i) - x_r(i)$ which is simplified to $(x_r^1(i), x_r^0(i))$ for matched data sets with a binary outcome. MF also defines a difference variable d_r^* for each exposure variable computed by Equation 6.2. This variable is the simplified version of the difference vector computed by Equation 6.11. When the outcome is binary, our generalized method creates a two-dimensional vector $d_r^* = (d_r^{*0}, d_r^{*1})$ for each exposure variable x_r from Equation 6.11 which is simplified for each individual variable d_r^{*0} and d_r^{*1} as

$$d_r^{*t}(i) = \begin{cases} x_r^t(i) - x_r^{(1-t)}(i) & \text{for } i \in \{1, 2, \ldots, N\}, \\ x_r^{(1-t)}(i-N) - x_r^t(i-N) & \text{for } i \in \{N+1, N+2, \ldots, 2N\} \end{cases} \tag{6.12}$$

for $t \in \{0, 1\}$ and $r \in \{1, 2, \ldots, R\}$. The data transformation method of MF uses only one of the variables d_r^{*0} and d_r^{*1}. These two variables are computed similarly and one of them is enough for identifying important variables.

To illustrate our method, consider Figure 6.3 which corresponds to a simulated matched data set with three outcome levels and one exposure variable x with no

effect. Figure 6.3a shows the scatter plot of x^0, x^1 and x^2 for each stratum in the original data set. The dashed line corresponds to the vector u. As can be seen in Figure 6.3a, all points are scattered randomly around the u and there is no region in this 3−dimensional space where x^0, x^1 and x^2 differ significantly from each other. Figure 6.3b shows the scatterplot of x^{*0}, x^{*1} and x^{*2} for this exposure variable in the transformed data set. The original strata are shown in blue (circle) and their counterfactuals are shown in red (×). The original strata and their counterfactuals are not clearly separated and a supervised learner will not classify well.

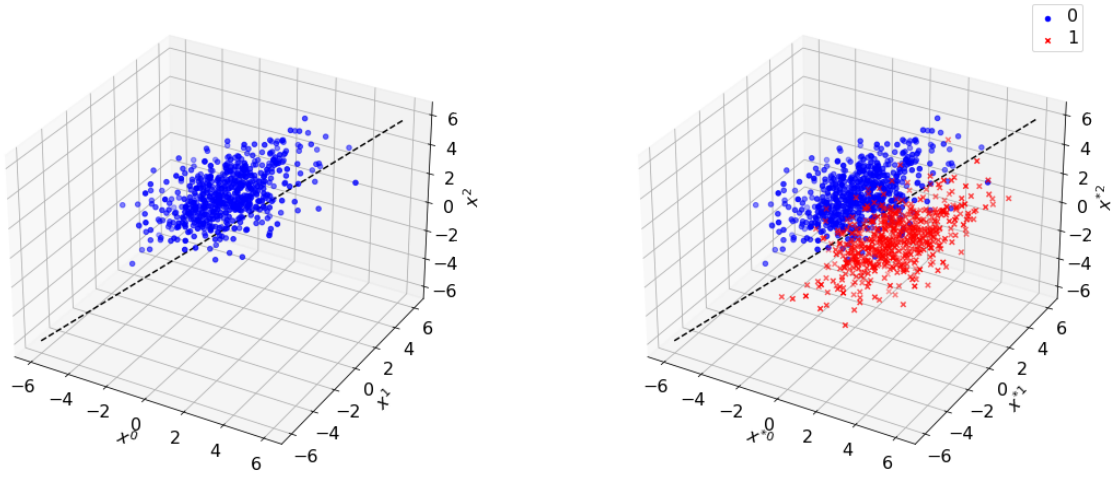


(a) (b)

Figure 6.3: The Scatter Plot of Exposure Variable x Simulated with No Effect for All Strata in Original (a) and Transformed Data Set (b).

Now, consider Figure 6.4 which corresponds to a matched data set with three outcome levels and an exposure variable x simulated with an effect. Figure 6.4a shows the original data set. For many of strata, there is a considerable difference between the values of x^0, x^1 and x^2. Figure 6.4b shows the scatterplot of x^{*0}, x^{*1} and x^{*2} for this exposure variable in the transformed data set. We can see in Figure 6.4b

that original strata and their counterfactuals are clearly separated and a supervised learner would classify well and identify the effect of this exposure variable.



(a) (b)

Figure 6.4: The Scatter Plot of Exposure Variable x Simulated with an Effect for All Strata in Original (a) and Transformed Data Set (b).

6.3.2 Variable Importance

Our algorithm applies a supervised learner on the transformed data set to identify important exposure and matching variables. The variable importance score of the supervised learner are modified to compute importance score for each matching and exposure variable in the matched data set.

To find important variables, we compute matched importance score denoted by MI for each exposure and matching variable. Variables with larger MI scores are more important. Let VI denote the original importance score obtained by the chosen classifier. The matched importance score of exposure variable x_r ($MI(x_r)$) is obtained

by the summation of the importance score of its associated variables as

$$MI(x_r) = \sum_{t=0}^{T-1} VI(x_r^{*t}) + \sum_{t=0}^{T-1} VI(d_r^{*t}) \tag{6.13}$$

for $r \in \{1, 2, \ldots, R\}$. If an exposure variable is important, we expect all of its associated variables help classify original strata and counterfactuals. Therefore, the summation is used to provide an overall score for the importance of x_r. When SHAP (Lundberg and Lee (2017) and Lundberg *et al.* (2018)) is used to measure the importance of each variable, the MI scores of each exposure variable is first computed for each stratum i and then to summarize these scores and obtain one measure for each variable, we take the mean magnitude of SHAP scores across all strata similar to Lundberg *et al.* (2018) and Lundberg *et al.* (2019). That is, we first compute the stratum level SHAP score for each exposure variable x_r as

$$MI(x_r(i)) = \sum_{t=0}^{T-1} VI(x_r^{*t}(i)) + \sum_{t=0}^{T-1} VI(d_r^{*t}(i)) \tag{6.14}$$

for $i \in \{1, 2, \ldots, 2N\}$ and $r \in \{1, 2, \ldots, R\}$. Then, the global SHAP score for x_r is computed as

$$MI(x_r) = \frac{1}{2N} \sum_{i=1}^{2N} |MI(x_r(i))| \tag{6.15}$$

for $r \in \{1, 2, \ldots, R\}$. As SHAP is an additive feature attribution method, Equation 6.14 gives the overall score of the group of variables including $\{x_r^{*0}, x_r^{*1}, \ldots, x_r^{*(T-1)}\}$ for stratum i.

The matched importance of a matching variable v_m ($MI(v_m)$) is also computed as

$$MI(v_m) = VI(v_m^+) \tag{6.16}$$

for $m \in \{1, 2, \ldots, M\}$. When SHAP is used to measure the importance of a variable, the matched importance score of the matching variable v_m for stratum i ($MI(v_m(i))$)

is first computed as

$$MI(v_m(i)) = VI(v_m^+(i)) \tag{6.17}$$

for $m \in \{1, 2, \ldots, M\}$ and $i \in \{1, 2, \ldots, 2N\}$. Then, its global score across all strata is computed as

$$MI(v_m) = \frac{1}{2N} \sum_{i=1}^{2N} |MI(v_m(i))| \tag{6.18}$$

for $m \in \{1, 2, \ldots, M\}$.

6.4 Experiments

In this section, we evaluate the performance of our proposed method through simulation studies and a real clinical application. Section 6.4.1 explains the data generation procedure, parameter settings of the Neural Network classifier and results for our simulations. Section 6.4.2 explains the usefulness of our method in a real clinical application. We did not compare our method with traditional methods for matched data sets in our simulations, because our experiments on Chapter 3 showed the limitation of these methods in identifying important variables from high-dimensional matched data sets with interaction effects. However, we compared the performance of our method with binary conditional logistic regression for the real data set which has small number of variables.

6.4.1 Simulation Study

We simulated matched data sets with nominal outcome with the number of levels ranging from 3 to 7. Each matched data set has 600 strata, 100 exposure variables and 5 matching variables. We consider a simple scenario with one important exposure variable and a complex scenario with an interaction effect between two exposure variables where the effect of one exposure variable depends on the value of the other expo-

sure. Each matching variable v_m is generated independently according to the Poisson distribution with the parameter equal to 5. The values of each exposure variable x_r under the outcome level t (x_r^t) is generated as $x_r^t = b_r + d_r^t$ where b_r, unless otherwise stated, is uniformly distributed between 1 and 50, and d_r^t is normally distributed with the mean of μ_r^t and the standard deviation of 1. In our simulations, to generate the exposure variable x_r with no effect, μ_r^t is set to 0 for all $t \in \{0, 1, \ldots, T-1\}$. Also, to generate the exposure variable x_r with an effect, $\mu_r^0 \in \{-2, -1.5, -1, 1, 1.5, 2\}$ and $\mu_r^t = 0$ for $t \in \{1, 2, \ldots, T-1\}$. We use $|\mu_r^0|$ as the measure of effect size. Larger values of $|\mu_r^0|$ indicate larger effect size. More details regarding data simulation for the simple and complex scenarios are provided in Sections 6.4.1 and 6.4.1 respectively.

For each simulation scenario, we generated 10 data sets to account for the randomness in our variables and used Neural Network as the classifier because it handles high-dimensionality, non-linear and interaction effects. The neural network architecture used in our simulations consists of 3 fully connected layers of size 30 followed by an output layer of size 2. To avoid overfitting due to the large number of variables, we used L1 regularization with penalty parameter equal to 0.008 for weights connecting the input layer to the first hidden layer. We used Adam optimization algorithm for training the Neural Network algorithm. Also, the learning rate is set to 0.001, number of epochs is set to 250, and batch size is set to 5. We used 5-fold cross validation to split 600 strata into training and test sets. The neural network algorithm is trained on the training set and the trained model is used to predict output for each stratum in the test set. We used DeepSHAP algorithm (Lundberg and Lee (2017)) to estimate SHAP scores of each variable using the predictions of Neural Network model for strata in the test set. DeepSHAP provides SHAP scores for each output node of the model. The Neural Network model trained on the transformed data set has two output nodes. The first and second output node of Neural Network

model take 0 and 1 for original strata and 1 and 0 for counterfactuals. When Neural Network model has two output nodes, the estimated SHAP scores for the two output nodes have similar value but different sign. In our simulations, we used SHAP scores corresponding to the first output node. The matched importance scores for exposure variables are computed based on Equations 6.14 and 6.15 and for matching variables based on Equations 6.17 and 6.18.

Simple Scenario: One Important Exposure Variable

Here, we simulated matched data sets with an effect for exposure variable x_1 and no effect for other exposure variables $\{x_2, x_3, \ldots, x_{100}\}$. That is, we simulated x_1 such that the distribution of x_1^t differ among the T outcome levels. In particular, we set $\mu_1^0 \in \{1, 1.5, 2\}$ and $\mu_1^t = 0$ for $t \in \{1, 2, \ldots, (T-1)\}$. For the remaining exposure variables with no effect, $\mu_r^t = 0$ for $r \in \{2, 3, \ldots, 100\}$ and $t \in \{0, 1, \ldots, T-1\}$.

Figure 6.5 shows MI scores for the top 20 exposure variables with largest MI scores for effect size $(|\mu_1^0|) \in \{1, 1.5, 2\}$ and $T \in \{3, 5, 7\}$. Each plot corresponds to a specific effect size $(|\mu_2^0|)$ and number of outcome levels (T). As can be seen in this Figure, the effect of exposure variable x_1 has been detected by our method because it received a significantly larger importance scores than other exposure variables in all of these 9 plots. The situation with effect size of 2 and $T = 3$ is an easy case which is shown at the top left corner of Figure 6.5. As the effect size becomes smaller for a fixed value of T (moving from top to bottom), identifying the effect of x_1 becomes harder because the difference between the importance of x_1 and other noise variables becomes smaller. The change in T for a fixed effect size (moving from left to right) does not change MI score of exposure variables significantly for this simple scenario, and they remain fairly consistent.

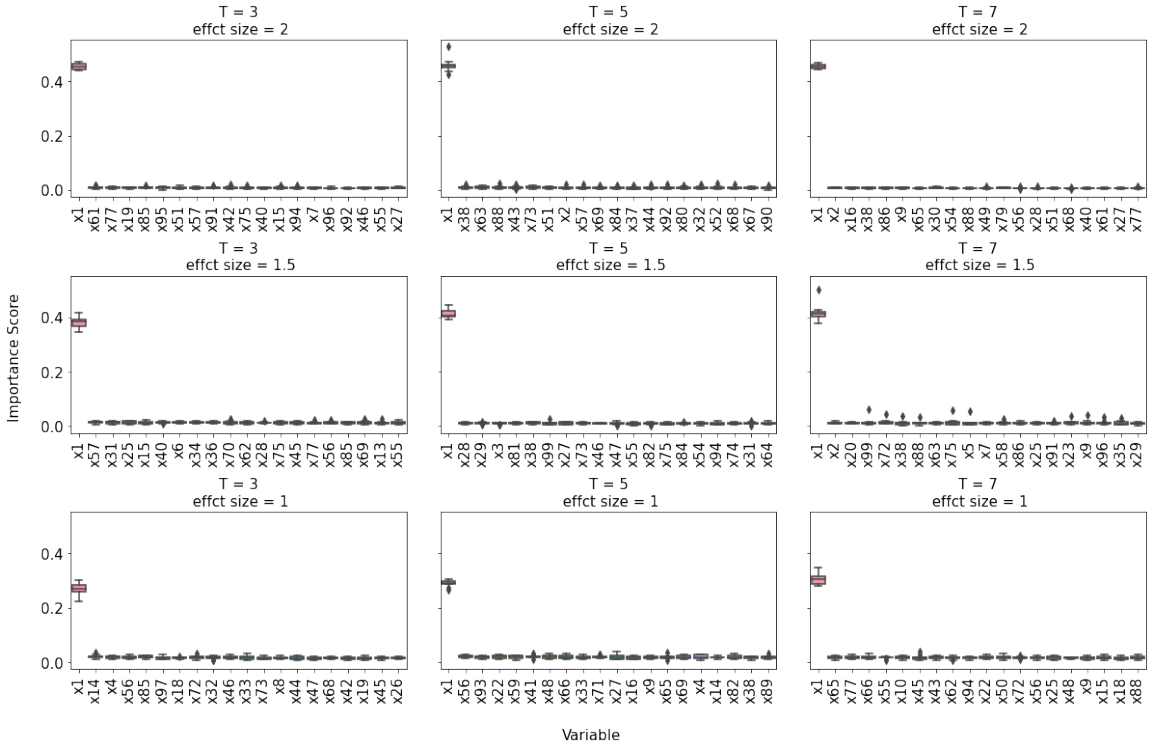


Figure 6.5: Simple Scenario with One Important Exposure Variable x_1: Matched Importance Score of the Top 20 Exposure Variables with Largest Scores for Effect Size ($|\mu_1^0|$) $\in \{1, 1.5, 2\}$ and $T \in \{3, 5, 7\}$.

Complex Scenario: Interaction Between Two Exposure Variables

Here, we consider a more complex scenario and simulated matched data sets with an interaction effect between the two exposure variables x_1 and x_2 and no effect for other exposure variables $\{x_3, x_4, \ldots, x_{100}\}$. In these simulated data sets, when exposure variable x_2 is considered individually, no effect is observed, however, when it is considered with exposure variable x_1, we observe an effect which changes with the values of x_1. In particular, when $b_1 < 25$, we set $\mu_2^0 \in \{1, 1.5, 2\}$ and when $b_1 \geq 25$, we set $\mu_2^0 \in \{-1, -1.5, -2\}$. For all other outcome levels $t \in \{1, 2, \ldots, T-1\}$, $\mu_2^t = 0$. The value for μ_2^0 is selected such that in each simulation, the average of d_2^0 over all

strata is equal to zero. This forces the variable x_2 to be important only when it is considered with exposure variable x_1. For the remaining exposure variables with no effect, $\mu_r^t = 0$ for $r \in \{3, 4, \ldots, 100\}$ and $t \in \{0, 1, \ldots, T-1\}$.

Figure 6.6 represents MI scores of the top 20 exposure variables with largest scores for effect size $(|\mu_2^0|) \in \{1, 1.5, 2\}$ and $T \in \{3, 5, 7\}$. Each plot corresponds to a specific effect size $(|\mu_2^0|)$ and number of outcome levels (T). We can see in nearly all of these plots that exposure variables x_1 and x_2 have received significantly larger MI scores than other noise variables. Thus, the effect of x_1 and x_2 have been detected by our method. The plot on the top left corresponding to $T = 3$ and effect size of 2 is the easy case and used as the reference. As the effect size decreases (moving from top to bottom) or T increases (moving from left to right), identifying the effects of x_1 and x_2 becomes harder, because their MI scores become closer to the MI scores of noise variables. We observed in our simulations in Chapter 3 that MF for binary outcome $(T = 2)$ performs better than CLR in identifying interaction effects. The performance of MML method for $T > 2$ is also as good as MF, and we see in Figure 6.6 that the effects of both exposure variables x_1 and x_2 are detected. Thus, we expect that MML also performs better than CLR in identifying interaction effects.

6.4.2 Clinical Application

We considered the data set EEG which can be accessed through R MANOVA.RM package. This data set contains information of 160 patients who were diagnosed with either Alzheimer disease (AD), mild cognitive impairment (MCI), or subjective cognitive complaints (SCC). For each patient, z-scores for brain rate and Hjorth complexity are measured at three brain regions including frontal, temporal, and central. The subject specific factors considered in this data set are sex (men vs. women), diagnosis (AD vs. MCI vs. SCC) and age (< 70 vs. ≥ 70). In addition, within subject factors

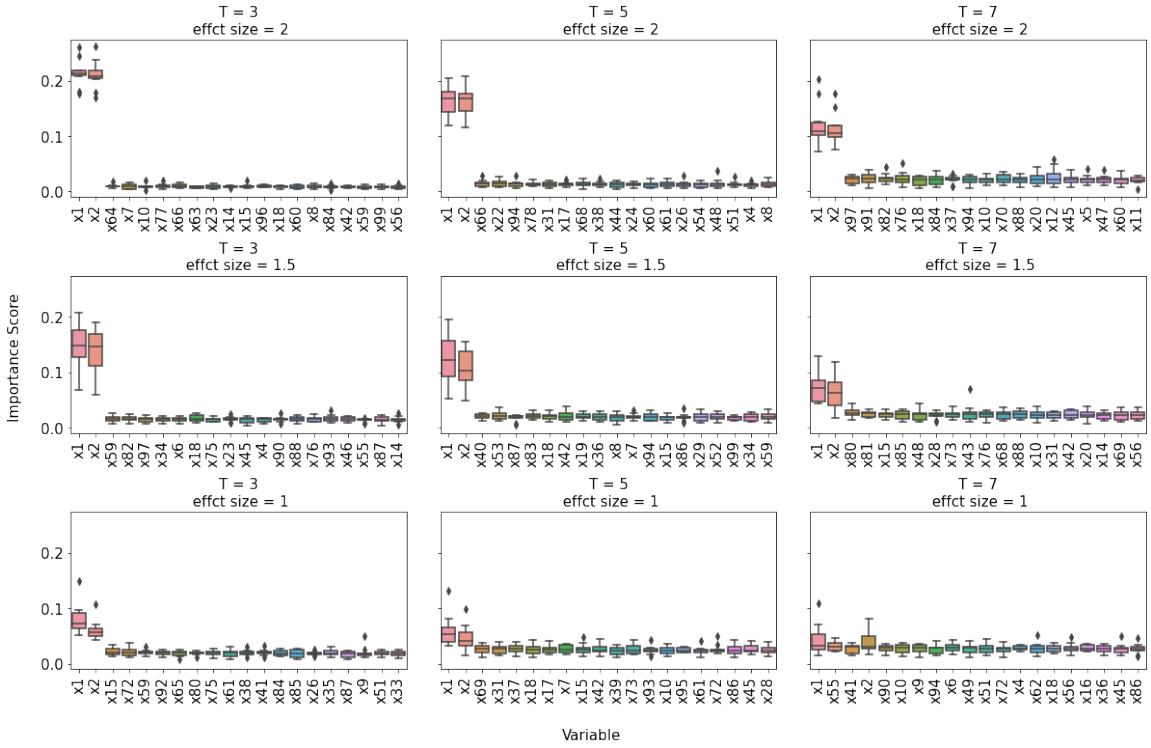


Figure 6.6: Complex Scenario with an Interaction Effect Between Two Exposure Variables x_1 and x_2: Matched Importance Scores of the Top 20 Exposure Variables with Largest Scores for Effect Size ($|\mu_2^0|$) $\in \{1, 1.5, 2\}$ and $T \in \{3, 5, 7\}$.

are brain region (frontal, temporal, central) and numerical EEG variables including brain rate and complexity. The research question here is whether the structure of EEG variables (brain rate and complexity) differ across three brain regions (frontal, temporal, central).

This data set has the structure of a matched study design. Each subject can be considered as a matched set or stratum and the three subject specific variables (sex, diagnosis and age) which are consistent within each stratum can be considered as matching variables. The two EEG features (brain rate and complexity) are to be compared across brain regions as two exposure variables, and the brain region is considered as the nominal outcome of interest with three categories or levels ($T = 3$).

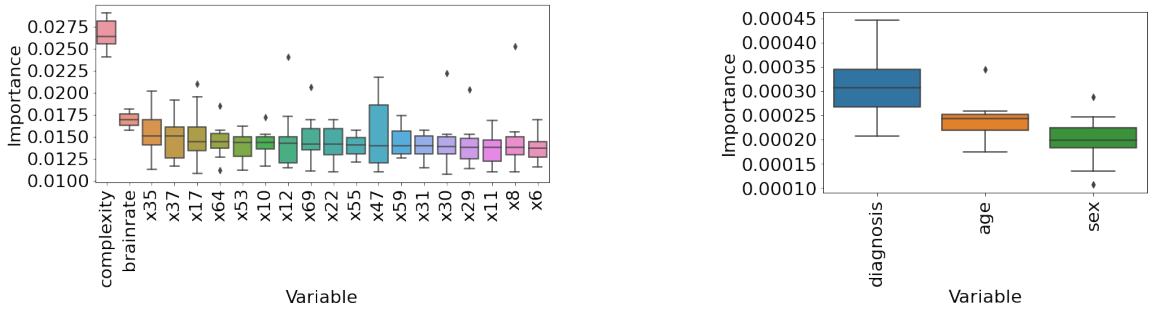
To evaluate the matched importance score of the two exposure variables including brain rate and complexity, we augmented this data set by including 70 other exposure variables simulated with no effect. The matched importance scores of these 70 simulated exposure variables is used to set a threshold to identify important EEG variables. We denote these simulated exposure variables by $\{x_1, x_2, \ldots, x_{70}\}$. The simulation procedure for variables $\{x_1, x_2, \ldots, x_{70}\}$ is similar to what was described in section 6.4 for exposure variables with no effect. We simulated these 70 exposure variables 10 times to account for the randomness in our simulated variables, applied our method on this data set and measured the MI score of each variable based on Equations 6.13 and 6.16 for exposure and matching variables, respectively.

We used Random Forest as the classifier because it handles high-dimensionality, mixed numerical and categorical variables, non-linear and interaction effects. The number of trees in Random Forest is set to 500, the number of variables selected at each split is set to $\sqrt{(}p)$ (p is the number of variables in the transformed data set), and trees were grown to purity. We also used 5 fold cross validation to separate data into training and test and used the average of variable importance scores over 5 folds as the importance score (VI) of each variable.

Figure 6.7 shows the MI scores of matching and exposure variables obtained from our method. Figure 6.7a and Figure 6.7b show respectively the MI scores of the 20 most important exposure variables and the MI scores of matching variables for 10 simulated data set. We can observe in Figure 6.7a that variable *complexity* has received considerably larger MI score than simulated variables with no effect. But, the MI score of exposure variable *brain rate* is not considerably larger than the simulated noise variables. Thus, we can conclude that the values of *complexity* differ among frontal, central and temporal brain regions in this data set. From Figure 6.7b,

we can observe that matching variable diagnosis has received larger MI scores than variables age and sex.

For comparison, Conditional Logistic Regression (CLR) model was applied for each pair of outcome levels. We used *central* outcome level as the reference and fit two CLR models where one compares the outcome level *frontal* with *central* and the other one compares the outcome level *temporal* with *central*. If the p-value of a variable obtained from any of these two models is significant at $\alpha = 0.05$, the variable is selected as important. Our results show that CLR did not select exposure variables *complexity* and *brain rate* in any of 10 simulated data sets, but it selected incorrectly some of the variables simulated with no effect. Therefore, the variables selected by our method and CLR are different. This difference can be due to existence of interactions between variables that CLR did not detect.



(a) (b)

Figure 6.7: EEG Data Set: Matched Importance Scores of Exposure (a) and Matching (b) Variables.

6.5 Conclusion

We proposed a new machine learning algorithm to identify important variables from high-dimensional matched data sets with more than two outcome levels. Our

method is suitable for matched data sets with hundreds or thousands of variables, and is developed to detect complex non-linear and interaction effects. We showed our method is the generalization of Matched Forest algorithm proposed for matched data sets with binary outcome. Our method first transforms matched data set to supervised setting and then applies a supervised learner with modified variable importance score on the transformed data set to identify important exposure and matching variables.

Our simulation studies showed the effectiveness of our method in identifying complex interactions among variables. Also, the analysis on the clinical data set showed that results from our method can be different from alternative approaches because of its ability to detect complex interaction effects.

Chapter 7

CONCLUSION

In this dissertation, machine learning models have been proposed for the analysis of high-dimensional matched data sets with hundreds or thousands of exposure variables and dozens of matching variables. The proposed models in this dissertation are designed for the task of variable selection, effect estimation and classification in high-dimensional matched data sets. The proposed methods are effective in high-dimensional settings where interaction among variables exists.

In Chapter 3, we proposed Matched Forest (MF) to identify important variables from high-dimensional matched case-control data sets. The outcome of interest is binary (case or control) and each stratum consists of one case and one control unit. Our experiments showed that MF is effective in identifying complex non-linear and interaction effects. This work is published in Shomal Zadeh *et al.* (2020).

In Chapter 4, we proposed three enhancements of Matched Forest (MF). First, we proposed Weighted Matched Forest (WMF) which adaptively regularizes MF to focus on highly important variables for splits. Our results showed that WMF has better variable selection performance than MF. Second, we generalized the application of MF to classification problems. we explained how MF is used to classify an unlabeled pair to either case-control or control-case. Our results showed that MF not only performs well in variable selection, but also has a better classification accuracy than existing algorithms. Finally, in the third enhancement, we proposed two new metrics to estimate the effect of variables identified as important by MF.

In Chapter 5, we generalized MF to matched case-control study designs where multiple controls are matched to each case. We also used Neural Network with SHAP

scores to identify important variables from high-dimensional matched case-control data sets. This method is referred to as Matched Neural Network (MNN), and our results showed that it performs better than MF for identifying interaction effects when number of strata is sufficiently large.

In Chapter 6, we generalized our variable selection method to matched data sets where outcome has more than two levels. Our method is referred to as Multinomial Matched Learner (MML) which aims at identifying important variables from high-dimensional matched data sets with multiple outcome levels. Our results showed the superiority of our method in identifying important variables compared with existing methods in the literature.

The application of our methods is not limited to only matched study designs. One research area which may benefit from our methods is causal inference. The traditional methods for estimation of causal effect of a treatment variable on the outcome focused on the population average treatment effect. However, the causal effect of a variable on the outcome may not be constant over the entire population and it may change depending on other variables. This concept is known as heterogeneous treatment effect or effect modification, and variables across which the causal effect of interest differs are referred as effect modifiers. Our proposed methods can be combined with outcome based modeling approaches to identify effect modifiers.

Adewale, A. J., I. Dinu and Y. Yasui, "Boosting for correlated binary classification", Journal of computational and graphical statistics **19**, 1, 140–153 (2010).

Ancona, M., E. Ceolini, C. Öztireli and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks", arXiv preprint arXiv:1711.06104 (2017).

Asafu-Adjei, J., T. Mahlet G., B. Coull, R. Balasubramanian, M. Lev, L. Schwamm and R. Betensky, "Bayesian variable selection methods for matched case-control studies", The International Journal of Biostatistics **13**, 1 (2017).

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation", PloS one **10**, 7, e0130140 (2015).

Balasubramanian, R., E. A. Houseman, B. A. Coull, M. H. Lev, L. H. Schwamm and R. A. Betensky, "Variable importance in matched case–control studies in settings of high dimensional data", Journal of the Royal Statistical Society: Series C (Applied Statistics) **63**, 4, 639–655 (2014).

Basu, S., K. Kumbier, J. B. Brown and B. Yu, "Iterative random forests to discover predictive and stable high-order interactions", Proceedings of the National Academy of Sciences **115**, 8, 1943–1948 (2018).

Becher, K.-H. and K.-H. Jöckel, "Bias adjustment with polychotomous logistic regression in matched case-control studies with two control groups", Biometrical journal **32**, 7, 801–816 (1990).

Bhojwani, D., H. Kang, N. P. Moskowitz, D.-J. Min, H. Lee, J. W. Potter, G. Davidson, C. L. Willman, M. J. Borowitz, I. Belitskaya-Levy *et al.*, "Biologic pathways associated with relapse in childhood acute lymphoblastic leukemia: a children's oncology group study", Blood **108**, 2, 711–717 (2006).

Breiman, L., "Random forests", Machine learning **45**, 1, 5–32 (2001).

Breiman, L., "Manual on setting up, using, and understanding random forests v3. 1", Statistics Department University of California Berkeley, CA, USA **1**, 58 (2002).

Breiman, L., J. Friedman, R. A. Olshen and C. J. Stone, *Classification and regression trees* (Chapman and HallCRC, 1984).

Breslow, N. E. and D. G. Clayton, "Approximate inference in generalized linear mixed models", Journal of the American statistical Association **88**, 421, 9–25 (1993).

Chen, H., S. Lundberg and S.-I. Lee, "Explaining models by propagating shapley values of local components", arXiv preprint arXiv:1911.11888 (2019).

Cox, D. R. and E. J. Snell, *Analysis of binary data*, vol. 32 (CRC Press, 1989).

Friedman, J. H., "Greedy function approximation: a gradient boosting machine", Annals of statistics pp. 1189–1232 (2001).

Gebregziabher, M., P. Guimaraes, W. Cozen and D. V. Conti, "A polytomous conditional likelihood approach for combining matched and unmatched case–control studies", Statistics in medicine **29**, 9, 1004–1013 (2010).

He, H., P. Wu and D.-G. Chen, *Statistical causal inferences and their applications in public health research* (Springer, 2016).

Heller, R., E. Manduchi and D. S. Small, "Matching methods for observational microarray studies", Bioinformatics **25**, 7, 904–909 (2008).

Ho, D. E., K. Imai, G. King and E. A. Stuart, "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference", Political analysis **15**, 3, 199–236 (2007).

Hosmer, D. W. and S. Lemeshow, *Applied logistic regression* (Wiley New York, 2000).

Hosmer Jr, D. W., S. Lemeshow and R. X. Sturdivant, *Applied logistic regression*, vol. 398 (John Wiley & Sons, 2013).

Keogh, G., "A matched pairs analysis of international protection outcomes in ireland", arXiv preprint arXiv:1705.10131 (2017).

Le Hesran, J.-Y., J. Akiana, E. H. M. Ndiaye, M. Dia, P. Senghor and L. Konate, "Severe malaria attack is associated with high prevalence of ascaris lumbricoides infection among children in rural senegal", Transactions of the Royal Society of Tropical Medicine and Hygiene **98**, 7, 397–399 (2004).

Liang, K.-Y. and W. F. Stewart, "Polychotomous logistic regression methods for matched case-control studies with multiple case or control groups", American journal of epidemiology **125**, 4, 720–730 (1987).

Liaw, A. and M. Wiener, "Classification and regression by randomforest", R News **2**, 3, 18–22, URL `https://CRAN.R-project.org/doc/Rnews/` (2002a).

Liaw, A. and M. Wiener, "Classification and regression by randomforest", R News **2**, 3, 18–22, URL `https://CRAN.R-project.org/doc/Rnews/` (2002b).

Lichman, M., "UCI machine learning repository", URL `http://archive.ics.uci.edu/ml` (2013).

Lundberg, S. and S.-I. Lee, "A unified approach to interpreting model predictions", arXiv preprint arXiv:1705.07874 (2017).

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal and S.-I. Lee, "Explainable ai for trees: From local explanations to global understanding", arXiv preprint arXiv:1905.04610 (2019).

Lundberg, S. M., G. G. Erion and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles", arXiv preprint arXiv:1802.03888 (2018).

McCulloch, C. E. and J. M. Neuhaus, "Generalized linear mixed models", Encyclopedia of biostatistics **4** (2005).

Molenberghs, G. and G. Verbeke, *Models for discrete longitudinal data* (Springer Science & Business Media, 2006).

Mukherjee, B., I. Liu and S. Sinha, "Analysis of matched case–control data with multiple ordered disease states: possible choices and comparisons", Statistics in Medicine **26**, 17, 3240–3257 (2007).

Neyman, J., "On the application of probability theory to agricultural experiments. essay on principles. section 9.", Statistical Science **5**, 465–480 (1923).

Peleg, A. Y., S. Husain, Z. A. Qureshi, F. P. Silveira, M. Sarumi, K. A. Shutt, E. J. Kwak and D. L. Paterson, "Risk factors, clinical characteristics, and outcome of nocardia infection in organ transplant recipients: a matched case-control study", Clinical Infectious Diseases **44**, 10, 1307–1314 (2007).

Qian, J., S. Payabvash, A. Kemmling, M. H. Lev, L. H. Schwamm and R. A. Betensky, "Variable selection and prediction using a nested, matched case-control study: Application to hospital acquired pneumonia in stroke patients", Biometrics **70**, 1, 153–163 (2014).

Rose, S. and M. J. Van der Laan, "Why match? investigating matched case-control study designs with causal effect estimation", The international journal of biostatistics **5**, 1 (2009).

Rose, S. and M. J. van der Laan, "Why Match? Investigating Matched Case-Control Study Designs with Causal Effect Estimations", The International Journal of Biostatistics **5**, 1, 1 (2009).

Rosenbaum, P. R., "Impact of multiple matched controls on design sensitivity in observational studies", Biometrics **69**, 1, 118–127 (2013).

Rothman, K. J., S. Greenland and T. L. Lash, *Modern epidemiology* (Lippincott Williams & Wilkins, 2008).

Rubin, D. B., "Estimating causal effects of treatments in randomized and nonrandomized studies.", Journal of educational Psychology **66**, 5, 688 (1974).

Rubin, D. B., "Assignment to treatment group on the basis of a covariate", Journal of educational Statistics **2**, 1, 1–26 (1977).

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization", in "Proceedings of the IEEE international conference on computer vision", pp. 618–626 (2017).

Shapley, L. S., "A value for n-person games", Contributions to the Theory of Games **2**, 28, 307–317 (1953).

Shomal Zadeh, N., S. Lin and G. C. Runger, "Matched forest: supervised learning for high-dimensional matched case–control studies", Bioinformatics **36**, 5, 1570–1576 (2020).

Shomalzadeh, N., S. Lin and G. C. Runger, "Matched forest: Supervised learning for high-dimensional matched case-control studies", Bioinformatics (2019).

Shrikumar, A., P. Greenside and A. Kundaje, "Learning important features through propagating activation differences", arXiv preprint arXiv:1704.02685 (2017).

Shrikumar, A., P. Greenside, A. Shcherbina and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences", arXiv preprint arXiv:1605.01713 (2016).

Simonyan, K., A. Vedaldi and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps", arXiv preprint arXiv:1312.6034 (2013).

Splawa-Neyman, J., D. M. Dabrowska and T. Speed, "On the application of probability theory to agricultural experiments. essay on principles. section 9.", Statistical Science pp. 465–472 (1990).

Springenberg, J. T., A. Dosovitskiy, T. Brox and M. Riedmiller, "Striving for simplicity: The all convolutional net", arXiv preprint arXiv:1412.6806 (2014).

Stanfill, B., S. Reehl, L. Bramer, E. S. Nakayasu, S. S. Rich, T. O. Metz, M. Rewers, B.-J. Webb-Robertson and T. S. Group, "Extending classification algorithms to case-control studies", Biomedical engineering and computational biology **10**, 1179597219858954 (2019).

Strobl, C., A.-L. Boulesteix, A. Zeileis and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution", BMC bioinformatics **8**, 1, 25 (2007).

Strobl, C. and A. Zeileis, "Danger: High power!–exploring the statistical properties of a test for random forest variable importance", (2008).

Stuart, E. A., "Matching methods for causal inference: A review and a look forward", Statistical science: a review journal of the Institute of Mathematical Statistics **25**, 1, 1 (2010).

Sundararajan, M., A. Taly and Q. Yan, "Gradients of counterfactuals", arXiv preprint arXiv:1611.02639 (2016).

Szyszkowicz, M., "Use of generalized linear mixed models to examine the association between air pollution and health outcomes", International journal of occupational medicine and environmental health **19**, 4, 224–227 (2006).

Tan, Q., M. Thomassen and T. A. Kruse, "Feature selection for predicting tumor metastases in microarray experiments using paired design", Cancer Informatics **3**, 213 (2007).

Tsou, J. A., J. S. Galler, K. D. Siegmund, P. W. Laird, S. Turla, W. Cozen, J. A. Hagen, M. N. Koss and I. A. Laird-Offringa, "Identification of a panel of sensitive and specific dna methylation markers for lung adenocarcinoma", Molecular cancer **6**, 1, 70 (2007).

Tutz, G. and H. Binder, "Generalized additive modeling with implicit variable selection by likelihood-based boosting", Biometrics **62**, 4, 961–971 (2006).

Vierkant, R. A., T. M. Therneau, J. L. Kosanke and J. M. Naessens, "A sas macro to analyze data from a matched or finely stratified case-control design", in "Proceedings of the 24th Annual SAS User's Group International Conference. SAS Institute, Inc., Miami Beach, FL", (1999).

Zeiler, M. D. and R. Fergus, "Visualizing and understanding convolutional networks", in "European conference on computer vision", pp. 818–833 (Springer, 2014).

Zhou, J. and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model", Nature methods **12**, 10, 931–934 (2015).

Zintgraf, L. M., T. S. Cohen, T. Adel and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis", arXiv preprint arXiv:1702.04595 (2017).