A Subgroup Identification Method

with Interaction Filtering and Quantitative Criteria

by

YAN SUN
Bachelor, Fudan University, Shanghai, China, 2007
Master of Science, Fudan University, Shanghai, China, 2010

Thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Chicago, 2015

Chicago, Illinois

Defense Committee:
Samad Hedayat, Chair and Advisor
Jing Wang
Jie Yang
Min Yang
Hua Yun Chen, School of Public Health

This thesis is dedicated to my wife Ling Cheng, my parents, and my parents in law.

ACKNOWLEDGMENT

I wish to express my sincere thanks to my advisor Professor Samad Hedayat for his extraordinary support and guidance of my Ph.D. study as well as my career building. His patience, enthusiasm and experience have profoundly influenced me and made me a better statistician.

I would like to thank Prof. Jing Wang, Prof. Jie Yang, Porf. Min Yang, Prof. Ryan Martin, Prof. Junhui Wang, and Prof. Hua Yun Chen for all the academic training and help I have received. Everything I learned from class has played an important role in my academic life and will continue to benefit me in my future career.

I would like to thank all my committee members, Prof. Samad Hedayat, Prof. Jing Wang, Prof. Jie Yang, Prof. Min Yang and Prof. Hua Yun Chen, for their kindness and help.

I would like to thank Dr. Viswanath Devanarayan and Dr. Xin Huang for their great help during my externship at Abbvie. The work they gave me was very interesting and inspiring.

I would like to thank my wife Ling Cheng for her great help and support, especially during the time I was writing my thesis. I would like to thank my parents Qi Jin and Ruzhong Sun for giving birth to me, raising me, educating me and supporting me for all these years.

Last, but not least, I want to thank everyone in my big family and all my friends. Being with them is a great pleasure of my life.

My research activity in preparing this dissertation is sponsored by the Abbvie Externship and also the United State National Science Foundation Award DMS-1306394 with Professor Samad Hedayat as its PI.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

SUMMARY

Subgroup identification has always been of great interest among the many functions and applications of statistical learning. In the pharmaceutical area, it is desirable to find a subgroup with enhanced treatment effect so that we can efficiently lower the number of patients required for a trail and improve the success rate of drug development projects. A more familiar name for this application is called personalized medicine, which has drawn great attention recently.

A majority of work has been done regarding the personalized medicine with their pros and cons. Some methods focus on the detection of subgroup effects but do not provide any way to select patients. Some methods have a tree regression style and provide a detailed picture of each patients performance, they are hence more optimized for prediction rather than subgroup identification. Some methods try to maximize the effect in the training dataset but tend to be too greedy. There are also methods trying to build a score system to stratify the patients.

In this dissertation, we propose a subgroup identification method with interaction filtering and quantitative criteria. More specifically, the method consists of two steps. Step 1 can select interaction covariates related to the individual treatment benefit without modeling the main effects. Step 2 can select a desired subgroup based on some quantitative criteria without relying on any specific model. The proposed method works for both the continuous and the survival response, and is shown to have a better performance than some popular existing methods.

CHAPTER 1

INTRODUCTION

1.1 Introduction to Subgroup Identification

With the era of big data coming, mining data through statistical learning algorithms for information that cannot be observed or understood directly has become more and more popular. Among the many functions and applications of statistical learning, subgroup identification has always been of great interest.

Subgroup identification can be broadly defined as a procedure that distinguishes a group of subjects from the rest of the population based on certain goals. For example, insurance companies may want to identify the people who are less likely to have an accident; shop owners may want to find out the customers that like coupons; universities may wish to identify the students that will succeed in the future. There could be numerous such goals and subgroups in different areas, and finding these subgroups will have a great impact in both industry and personal life.

Although it is relatively easy to specify the desired subgroup, how to actually find those subjects is often not obvious. In most situations, whether a subject belongs to the desired group is not observable at all or only observable in the collected datasets. For example, in ordinary clinical trials, it is usually not directly observable whether a specific patient can benefit much more from taking the treatment than taking the control, since patients can only be assigned to

one arm. Adverse events are only observable in the collected datasets, and what will happen to a new patient is not obvious. There are also cases that the grouping of a subject depends on others, such as the insurance company example in the previous paragraph. All these situations require statistical learning algorithms to construct some identification rule according to the goal so that whether a subject belongs to the desired subgroup can be decided. And of course, a training dataset is needed to perform this learning procedure.

The constructed identification rule that can be practically used to select subjects for a subgroup usually consist of part of the covariates or predictors in the training dataset, and the goal of the subgroup identification procedure is generally related to the response variable. The identification rule is constructed based on the relationship between the covariates and the response. Depending on the algorithms used, the identification rule could have different forms, such as a classification tree (1) or a polynomial. A tree-styled rule is essentially a series of yes/no questions that leads to a final subgroup, where the question is mostly about whether the value of a covariate is beyond a certain threshold. A polynomial usually produces a value that needs to be further associated with the final subgroup. But no matter what the identification rule looks like, one important problem is that not every covariate in the dataset is needed. Actually, only a small proportion of covariates are important in most situations. This is especially true when datasets have a large number of covariates, such as genomic data (2). Variable selection or dimension reduction methods are needed to handle the problem, and there are many such methods readily available. Traditionally for regression, we have naive ways like selecting variables based on p values of coefficients and best-subset selection (3),

neither of which performs well when there are more than a few covariates. Modern methods like forward/backward selection (4), principal component regression, LASSO (5) and other shrinkage methods are widely used.

Subgroup identification rule construction can lead to two possible results, the rule can be constructed and the rule cannot be constructed. There could be several reasons why the rule cannot be constructed. It is most likely that the method used is not powerful enough to detect the subgroup. Or it could be because none of the available covariates are relevant. In some cases, it is also possible that the desired subgroup does not exist, especially when the subjects' groups are not directly observable. Unfortunately, it is difficult to know which is the actual reason. When the rule can be constructed, it is still not the happy ending. It is entirely possible that the identified subgroup is a false discovery. A separate testing set or some validation procedure such as 5-fold cross-validation (6) are usually needed to validate the performance.

1.2 Subgroup Identification in the Pharmaceutical Area

One of the areas in which subgroup identification plays an important role is the pharmaceutical drug development. Inclusion and exclusion criteria commonly seen in a clinical trial is one basic form of subgroup selection, which helps increase the likelihood of producing reliable results. By selecting and examining a tailored subgroup, not only will the attrition rate of drug development projects be reduced, but we can also have a better understanding of the underlying relationship between the baseline covariates and the outcome of interest.

There are generally two types of subgroup identification in the pharmaceutical area. One is called the prognostic case, and the other is the predictive case. In the prognostic case, a

patients response irrespective of the treatment is of direct interest. For example, researchers may want to identify a subgroup of patients who will have a survival time less than three years. Or they may want to identify the subjects who will develop a certain symptom in the near future. In the predictive case, however, the focus is on the difference between outcome before and after the treatment. In other words, unlike the prognostic case, how long a patient can survive does not matter in a predictive case, but how much longer the patient can survive after taking the new treatment is of direct interest.

Use X to denote the patient's available covariates including an intercept covariate $\mathsf{1}$, Y to denote the patient's response, $\mathsf{T} = \pm\mathsf{1}$ to denote the patient's receiving the treatment and the control respectively. If we assume that the available covariates could well represent the patient, then what the prognostic case focuses on is $\mathsf{E}\big(\mathsf{Y} \mid \mathsf{X}\big)$, while the predictive case focuses on $\mathsf{E}\big(\mathsf{Y} \mid \mathsf{X}, \mathsf{T} = \mathsf{1}\big) - \mathsf{E}\big(\mathsf{Y} \mid \mathsf{X}, \mathsf{T} = -\mathsf{1}\big)$. In a linear regression setup, the following two models can be used to handle the two cases:

$$\mathsf{Y} = \beta'\mathsf{X} + \epsilon \tag{1.1}$$

$$\mathsf{Y} = \beta'\mathsf{X} + \gamma'\mathsf{X} \cdot \mathsf{T} + \epsilon \tag{1.2}$$

where β and γ are the coefficients of the covariate terms and the interaction terms respectively, and ϵ is the random noise. Note that interaction terms here also include the population treatment effect, which is expressed by the interaction between the treatment and the intercept covariate $\mathsf{1}$. Clearly, Model (1.1) can be used to handle the prognostic case, in which the treatment is not involved, and the direct relationship between the response and the covari-

ates is examined. In Model (1.2), it is easy to see that the interaction terms correspond to $\mathsf{E}(Y \mid X, T = 1) - \mathsf{E}(Y \mid X, T = -1)$, and thus the predictive case can be studied through estimating the interaction terms. Models (1.1) and (1.2) have a continuous response, but they can be easily extended to the binary and survival response by simply replacing Y with $\mathtt{logit}(Y)$and log hazard ratio respectively.

The covariates X in the above two models can be a wide variety of measurements, such as demographic information, clinical information or even genomic information. The term biomarker is frequently used in the literature in a very general sense to denote these measurements. In the prognostic case, the measurements that are relevant to the outcome of interest are called the prognostic markers. Similarly, in the predictive case, those measurements are called the predictive markers (7). The derived subgroup identification rule is sometimes called a signature. The population that satisfies the rule is called the signature-positive group, while the population that does not satisfy the rule is called the signature-negative group (8). A signature is also referred to as a predictive signature or a prognostic signature depending on the type of the subgroup identification procedure.

1.3 Personalized Medicine

A more familiar name for the predictive case of subgroup identification in the pharmaceutical area is personalized medicine, which has drawn a great deal of attention recently. President Obama mentioned in his State of the Union speech that his administration would launch "a new precision medicine initiative to bring us closer to curing diseases like cancer and diabetes". And it is widely believed that personalized medicine is the future of drug development.

The term personalized medicine is a little misleading, since it gives people the impression that the drug will soon be developed and customized at individual level. This might become true one day in the distant future. At present, however, personalized medicine mainly focuses on selecting patients for a treatment or selecting a treatment for a patient.

Productivity is currently one of the biggest problems in the pharmaceutical industry. Submission rate of new medicines and the success rate in clinical phases are decreasing, although the pharmaceutical research and development investment has increased significantly (9). The primary goal of a traditional clinical trial is usually to determine whether a new treatment is effective at the population level. However, people are heterogeneous in many ways. Finding a new therapy that works on average for the entire population is becoming more and more difficult, especially in the oncology area, where the same cancer can result from different reasons . For example, in a Phase III clinical trial comparing gefitinib with carboplatin plus paclitaxel as a first-line treatment for pulmonary adenocarcinoma, it was found out that only patients with epidermal growth factor receptor mutation had significantly longer progression-free survival time after receiving gefitinib (hazard ratio, 0.48; $p < 0.001$). But for patients lacking this mutation, progression free survival was significantly shorter in the gefitinib arm compared with the carboplatin plus paclitaxel arm (10). In cases like this, if the effectiveness is compared on the population level, it will not be surprising to see the trial fail. By finding a subgroup that is more likely to benefit from the treatment, we can efficiently reduce the variance, lower the number of patients required for a trail, improve the success rate of drug development projects

and finally control the ever-rising costs. The subgroup identification procedure designed for the above purpose can be considered as selecting patients for drugs.

On the other hand, even if the medicine can be shown effective on the population level, it does not mean that everyone in the population should take it. There almost always exists a subgroup of patients who cannot benefit from the treatment. Moreover, drug A being effective for the patient does not rule out the possibility that there could be a better drug B which has either a better efficacy or a lower toxicity for this patient. It is desirable to improve the quality of prescribing so that patients can take the medicine that is truly effective for them, and a more appropriate one if multiple options are available. The subgroup identification procedure designed for this purpose can be considered as selecting drugs for patients.

Though there are several successful examples of personalized medicine (11) (12) (13), the data-driven subgroup identification procedure, like any statistical learning activity, is only exploratory and suffers from false discovery. Subsequent confirmatory analysis that evaluates a small number of completely specified subgroups in clinical trials is often required (14) (15). When the discovered biomarkers do not work as expected, they will have a negative impact on the clinical trials, such as increasing the cost and prolonging the duration. Thus, having a good strategy of integrating biomarkers into clinical trials is very important (16), and FDA has released a draft guidance document (17).

1.4 Literatures on Methods for Personalized Medicine

The heterogeneity of treatment effects among patients can be explained by the interaction between the treatment and the covariates. Model (1.1) and (1.2) are two examples under the

linear regression setting. Hence studying the interactions is crucial to personalized medicine and is the main focus of many methods.

Traditionally, when performing exploratory subgroup analysis for the predictive case, treatment effects are often evaluated in several simple pre-specified subgroups based on some dichotomized covariates after the main analysis. This kind of approach can only handle simple interaction terms which are readily at hand, and often suffers from false discovery due to multiplicity issue. Many new methods are derived to better cope with the problem.

Bonetti and Gelber (18) invented a graphical technique called subpopulation treatment effect pattern plot (STEPP) to explore the existence of interaction effects among continuous covariates. Sauerbrei et al. (19) proposed an algorithm for multivariate model building with fractional polynomial interaction (MFPI) and compared it with STEPP. Bayman et al. (20) suggested a Bayesian approach using a hierarchical model with exchangeable mean responses between subgroups to detect potential qualitative interactions from a fixed number of predefined subgroups. Sivaganesan et al. (21) also introduced a Bayesian method which defines a separate class of models for every candidate predictor and uses a threshold on posterior model probabilities to determine the existence of subgroup effects. These methods mainly focus on the detection of interaction effects. But how a subgroup should be selected is not clear. Besides, they all lack the ability to handle the dataset with a large number of covariates.

Many subgroup identification methods for personalized medicine have a tree-style. Loh (22) developed an algorithm for regression tree construction, which has the ability of detecting interaction and eliminating variable selection bias. Foster et al. (23) introduced a two-step

method which first predicts the response difference between treatment and control for every patient and then uses this difference as the outcome in a regression or classification tree. These methods try to provide a detailed picture of how subjects perform in different terminal nodes of the tree. Hence the procedures are more optimized for response prediction rather than subgroup selection. They require a further strategy to combine multiple terminal nodes into subgroups, and have trouble in interpretation when terminal nodes consist of very different covariates.

Su et al. (24) proposed an interaction tree procedure for censored survival data. The method recursively divides the data into two subsets that have the largest interaction with the treatment, and a merging scheme is provided to combine terminal nodes into subgroups. Su et al. (25) later extended the algorithm to the continuous response. Lipkovich et al. (26) developed a subgroup identification method based on differential effect search (SIDES), which adopts the splitting idea from the interaction tree method, but with a different search strategy, and produces a list of promising candidate subgroups rather than merging different nodes. Chen et al. (8) modified the Patient Rule Induction Method (PRIM) (27) to perform subgroup identification. The method conducts a step-wise searching procedure through maximizing the treatment effect in the positive group while maintaining the interaction effect restriction. These methods aim at identifying subgroups with some maximized effect, which are more suitable for the purpose than those focusing on the prediction, but tend to be greedy sometimes. The interaction tree method also suffers from the interpretation problem when nodes with very different covariates are combined. SIZES and PRIM method have a fast decreasing sample size of the selected subgroup when more binary rules are added. Moreover, all the tree-style

methods that have been mentioned above lack the ability of handling the dataset with a large number of covariates.

Some methods try to identify patients based on some score or index. Zhao et al. (28) proposed to create a parametric scoring system using available covariates and associate the score with the treatment difference. Patients with scores above a pre-specified threshold will be in the desired group. Tian and Tibshirani (29) introduced an efficient algorithm to quickly select and dichotomize important covariates, and use the sum of the dichotomized covariates as an index score to stratify the population according to the individual treatment effect. The success of these methods relies on the quality of the score system.

There are two particularly interesting papers recently. Tian et al. (30) developed a simple method of modeling the interactions in a randomized clinical trial setup. It can detect the important interactions from a large set of biomarkers and can be flexibly coupled with different variable selection methods. The method utilizes the randomness of the treatment assignment and modifies the covariates so that the main effects are eliminated. A simple intuitive example is as follows. Let T be the treatment variable with $P(T = 1) = P(T = -1) = 0.5$ for every patient, where $T = 1$ and $T = -1$ correspond to the treatment and control arm respectively. Let Y be the response and suppose

$$Y = \beta' X + \gamma' X \cdot T + \epsilon \tag{1.3}$$

If we multiply Y by T, then we have

$$E(Y \cdot T \mid X) = E(\beta' X \cdot T \mid X) + E(\gamma' X \cdot T^2 \mid X) = \gamma' X$$

where the main effect is canceled out. One weak point of this method is that it only applies to the trial with balanced arms, i.e., two arms have the same number of patients. Another very interesting paper (31) looks at subgroup identification from a different point of view. While most other methods focus on the interaction and tries to select patients for a drug, this method is more about selecting drugs for patients and tries to maximizes the expected reward given that the treatment assignment rule is implemented, i.e., $\max\{E[Y \mid T = D(X)]\}$, where D is the treatment allocation rule based on the covariates. The method is called outcome weighted learning (OWL) and can be solved under the support vector machine framework.

1.5 Primary Interest of the Dissertation

This dissertation focuses on the subgroup identification methods for the predictive case, i.e., personalized medicine. A complete subgroup identification procedure usually includes the following steps:

a. Specify the feature or goal of the desired subgroup.

b. Select relevant predictors from a large collection of available candidate covariates.

c. Use a statistical learning algorithm to build the identification rule or say signature for the desired subgroup.

d. Evaluate the performance of the learning result.

In this dissertation, we propose a subgroup identification method consisting of step b and step c. The proposed method works for both the continuous and the survival response.

The introduced procedure for the variable selection step is inspired by Tian's work (30). It extends the original method to the case of unbalanced arms, i.e., two arms have different number of patients. The method focuses directly on the interaction terms, and the procedure is hence called interaction filtering.

It has been discussed that some subgroup identification methods are more optimized for prediction, which has the advantage of providing quantitative information when selecting patients, while some other methods tries to maximize the interaction effect, which focus more on selection rather than prediction. Our proposed method for the signature construction step tries to integrate the quantitative information into the subgroup identification procedure such that we can ask for a subgroup with a specified quantitative feature.

CHAPTER 2

INTERACTION FILTERING

In this chapter, we introduce an interaction filtering procedure that can select relevant interactions without modeling the main effects. The proposed method works for both the continuous and the survival response, and it is shown to be a reasonable procedure not relying on model specifications. We also conduct simulation to investigate the performance of the proposed method.

2.1 Motivation

The response of a patient depends on many things, such as the average treatment (or placebo) effect at the population level, the interaction between the treatment (or placebo) and the covariates, and the main effects of the covariates which does not depend on the therapy received. The average treatment (or placebo) effect at the population level can be viewed as the interaction between the treatment (or placebo) and the constant covariate 1. For simplicity, we use interaction to denote both the actual interaction and the average treatment (or placebo) effect at the population level in this dissertation.

Most of the popular variable selection methods are model-based. When constructing a model for the response, all relevant variables and terms have to be included. But in the field of personalized medicine, interaction is the focal point. It is desirable to construct a model with

only the interaction terms such that we do not have to model the main effects, and the variable selection procedure can solely focus on filtering the interactions.

2.2 Setup and Notations

Suppose we have a randomized clinical trial with two arms, in which patients are allocated at random to receive either the treatment or the control with some prespecified probabilities. The control could be a placebo or some other existing treatment. Let $Y \in R^1$ be the response variable and $T \in R^1$ be the binary treatment variable, where $T = 1$ and -1 correspond to the treatment arm and the control arm respectively. Let $X \in R^p$ be the p-dimensional random vector of baseline covariates, including an intercept covariate 1. Since it is a randomized trial, T and X are independent. Assume the observed dataset consists of N patients, and every patient data in the dataset is an independent and identically distributed copy of (Y, T, X).

2.3 Interaction Filtering for the Continuous Response

2.3.1 An Intuitive Derivation

When Y is a continuous response, the following multivariate regression model can usually be set up

$$Y = \beta_0' X + \gamma_0' XT + \epsilon \tag{2.1}$$

where β_0 and γ_0 are the coefficients of the main effect and the interaction effect respectively, and ϵ is the random error with mean 0 and variance σ^2. Note that every patient data is an independent and identically distributed copy of (Y, T, X).

We want to modify the distribution of Y given X, so that the main effects can be cancelled out on average. In order to do this, we construct a discrete random variable δ with the following support

$$\delta = \begin{cases} \alpha \text{ or } a\alpha & \text{when } T = 1; \\ \alpha \text{ or } b\alpha & \text{when } T = -1. \end{cases} \qquad a, b > 0 \text{ and } \neq 1 \tag{2.2}$$

such that

$$E\big(T\delta Y \mid X\big) = 2\alpha c\gamma_0' X \tag{2.3}$$

where $c > 0$ is a constant, and $\alpha > 0$ is an integer only for illustration purpose.

From (2.1) and (2.3), δ has to satisfy the following two conditions:

$$\begin{cases} E\big(\delta \mid T = 1\big) \cdot P\big(T = 1\big) = E\big(\delta \mid T = -1\big) \cdot P\big(T = -1\big) \\ E\big(\delta \mid T = 1\big) \cdot P\left(T = 1\right) + E\left(\delta \mid T = -1\right) \cdot P\left(T = -1\right) = 2c\alpha \end{cases} \tag{2.4}$$

where $P\left(T = 1\right)$ and $P\left(T = -1\right)$ are known after a trial is designed. Hence,

$$E\big(\delta \mid T = 1\big) = \frac{c\alpha}{P\left(T = 1\right)}, \quad E\big(\delta \mid T = -1\big) = \frac{c\alpha}{P\left(T = -1\right)}$$

Combining the above expectations with (2.2), we can then obtain the distribution of δ, which is as follows:

$$\begin{cases} P\big(\delta = \alpha \mid T = 1\big) = \frac{c - aP(T=1)}{(1-a)P(T=1)} \\ P\big(\delta = a\alpha \mid T = 1\big) = \frac{P(T=1) - c}{(1-a)P(T=1)} \\ P\big(\delta = \alpha \mid T = -1\big) = \frac{c - bP(T=-1)}{(1-b)P(T=-1)} \\ P\big(\delta = b\alpha \mid T = -1\big) = \frac{P(T=-1) - c}{(1-b)P(T=-1)} \end{cases} \tag{2.5}$$

With a proper choice of a, b and c, δ will be a random variable, i.e. values in (2.5) are all between 0 and 1, such that

$$E\big(\delta TY \mid X\big) = 2c\alpha\gamma_0' X$$

We estimate γ_0 through the ordinary least squares method. To make the new model more extensible and user friendly, we rewrite the least squares estimation process as follows:

$$\begin{aligned} & \min_{\gamma} \left\{ \sum_{i=1}^{N} \left(\delta_i T_i Y_i - 2c\alpha\gamma' X_i\right)^2 \right\} \\ \Longleftrightarrow \quad & \min_{\gamma} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{\delta_i^2} \left(T_i Y_i - 2c\alpha\gamma' X_i/\delta_i\right)^2 \right\} \\ & \alpha \text{ is selected after a proper choice of a, b, s.t. } \alpha, a\alpha, \text{ and } b\alpha \text{ are all integers} \\ \Longleftrightarrow \quad & \min_{\gamma} \left\{ \sum_{i=1}^{N} \sum_{j=1}^{\delta_i^2} \left(Y_i - 2c\alpha\gamma' X_i T_i/\delta_i\right)^2 \right\} \end{aligned} \tag{2.6}$$

The objective function (2.6) leads to the following interaction filtering procedure for the continuous response, which consists of a working dataset and a working model.

i. For patient i from 1 to N, replicate the corresponding observation δ_i^2 times, where δ_i is generated from the distribution of δ. The generated new dataset is called the working dataset.

ii. Use the working dataset from step i and the working model below to perform variable selection with an appropriate method such as LASSO under the context of the least squares procedure. The selected variables are then the ones that potentially interact with the treatment.

$$Y_i = 2c\alpha\gamma' X_i T_i/\delta_i + \epsilon_i, \text{ where } \epsilon \sim (0, \sigma^{*2}) \tag{2.7}$$

2.3.2 Justification of the Objective Function

The objective function (2.6), which estimates γ_0 using the working dataset and the working model, can be rewritten in the following equivalent general form

$$\min_f E_{\delta,T,Y}\left[\delta^2\left(Y - c\alpha f(X)\, T/\delta\right)^2 \mid X\right] \tag{2.8}$$
$$\text{for any given } X \text{ subject to } f \in \mathcal{F} = \{2\gamma' X \mid \gamma \in R^p\}$$

where the expectation is with respect to δ, T, Y given X, and $\mathcal{F}$ is a functional space. Let $f_0(X)$ and $f^*(X)$ be the respective minimizer of (2.8) when no specific form of $f(X)$ is assumed and when $f(X)$ is in the space of $\mathcal{F}$. We can show that there is a direct association between the individual treatment effect $E(Y \mid X, T = 1) - E(Y \mid X, T = -1)$ and $f_0(X)$ regardless of the actual relationship between Y and X. Hence, considering that $f^*(X)$ is a surrogate of $f_0(X)$

in a functional space, it is reasonable to use $f^*(X)$ for interaction filtering even if the model assumption in (2.1) is not fully correct.

The association between the individual treatment effect and $f_0(X)$ can be derived as follows.

$$\begin{aligned} & E_{\delta,T,Y}\left[\delta^2\left(Y - c\alpha f(X)\, T/\delta\right)^2 \mid X\right] \\ = \;& E\left[\delta^2 Y^2 - 2c\alpha f(X)\, T\delta Y + c^2\alpha^2 f^2(X)\, T^2 \mid X\right]. \end{aligned}$$

Define $L = \delta^2 Y^2 - 2c\alpha f T\delta Y + c^2\alpha^2 f^2 T^2$. Since δ and Y are independent when T is given, we have

$$\begin{aligned} E(L \mid X, T=1) =& E\left(\delta^2 \mid T=1\right) E\left(Y^2 \mid X, T=1\right) \\ & - 2c\alpha f(X)\, \frac{c\alpha}{P(T=1)} E(Y \mid X, T=1) + c^2\alpha^2 f^2(X) \end{aligned}$$

$$\begin{aligned} E(L \mid X, T=-1) =& E\left(\delta^2 \mid T=-1\right) E\left(Y^2 \mid X, T=-1\right) \\ & + 2c\alpha f(X)\, \frac{c\alpha}{P(T=-1)} E(Y \mid X, T=-1) + c^2\alpha^2 f^2(X). \end{aligned}$$

Hence,

$$\begin{aligned} & \min_f E(L \mid X) \\ \Leftrightarrow \;& \min_f \left\{ -2c^2\alpha^2 f(X)\, E(Y \mid X, T=1) + 2c^2\alpha^2 f(X)\, E(Y \mid X, T=-1) + c^2\alpha^2 f^2(X) \right\} \\ \Leftrightarrow \;& \min_f \left\{ -2f(X)\, E(Y \mid X, T=1) + 2f(X)\, E(Y \mid X, T=-1) + f^2(X) \right\} \end{aligned}$$

and we have the minimizer $f_0(X) = E(Y \mid X, T=1) - E(Y \mid X, T=-1)$. The minimizer $f_0(X)$ equals the individual treatment effect regardless of the actual relationship between Y and X.

2.3.3 Choice of Parameter a, b and c in general

It is sensible to use the working dataset, the working model and the least squares method to estimate γ_0 whether the assumed model is fully correct or not. Note that the least squares estimation procedure can also be written as

$$\min_{r} \sum_{i=1}^{N} \left(\frac{\delta_i T_i Y_i}{c\alpha} - 2\gamma' X_i \right)^2$$

thus one reasonable criterion for choosing a, b and c is to minimize the variance of $\frac{\delta TY}{c\alpha}$ for any given X. The variance can be derived as follows:

$$\begin{aligned}
& Var\left(\delta TY/c\alpha \mid X\right) \\
= \; & \frac{1}{c^2}\left[E\left(\delta^2 T^2 Y^2/\alpha^2 \mid X\right) - E^2\left(\delta TY/\alpha \mid X\right)\right] \\
= \; & \frac{1}{c^2}\left[E\left(\delta^2 Y^2/\alpha^2 \mid X\right) - E^2\left(\delta TY/\alpha \mid X\right)\right] \\
& \delta \perp\!\!\!\perp Y \mid X, T \\
= \; & \frac{1}{c^2}\left\{E_T\left[E\left(\delta^2/\alpha^2 \mid T\right) \cdot E\left(Y^2 \mid X, T\right)\right] - E_T^2\left[T \cdot E\left(\delta/\alpha \mid T\right) \cdot E\left(Y \mid X, T\right)\right]\right\}.
\end{aligned}$$

Note that

$$E\left(\delta^2/\alpha^2 \mid T = 1\right) = \frac{c - aP\left(T = 1\right)}{\left(1 - a\right)P\left(T = 1\right)} + a^2 \frac{P\left(T = 1\right) - c}{\left(1 - a\right)P\left(T = 1\right)} = \frac{c + ca - aP\left(T = 1\right)}{P\left(T = 1\right)},$$

$$E\left(\delta^2/\alpha^2 \mid T = -1\right) = \frac{c - bP\left(T = -1\right)}{\left(1 - b\right)P\left(T = -1\right)} + b^2 \frac{P\left(T = -1\right) - c}{\left(1 - b\right)P\left(T = -1\right)} = \frac{c + cb - bP\left(T = -1\right)}{P\left(T = -1\right)},$$

$$T \cdot E\big(\delta/\alpha \mid T=1\big) \cdot E\big(Y \mid X, T=1\big) = \frac{c}{P\,(T=1)} E\big(Y \mid X, T=1\big),$$

$$T \cdot E\big(\delta/\alpha \mid T=-1\big) \cdot E\big(Y \mid X, T=-1\big) = -\frac{c}{P\,(T=-1)} E\big(Y \mid X, T=-1\big).$$

Hence, we have

$$\begin{aligned} Var\big(\delta TY/c\alpha \mid X\big) \;=\; & \frac{c + ca - aP\,(T=1)}{c^2} E\left(Y^2 \mid X, T=1\right) \\ & + \frac{c + cb - bP\,(T=-1)}{c^2} E\left(Y^2 \mid X, T=-1\right) \\ & - \big[E\big(Y \mid X, T=1\big) - E\big(Y \mid X, T=-1\big)\big]^2. \end{aligned} \tag{2.9}$$

Next, we find the optimal a, b and c such that (2.9) is minimized for any given X. After taking derivative with respect to a and b, we have

$$\frac{\partial Var\big(\delta TY/c\alpha \mid X\big)}{\partial a} = \frac{E\big(Y^2 \mid X, T=1\big)}{c^2} \cdot \big[c - P\,(T=1)\big],$$

$$\frac{\partial Var\big(\delta TY/c\alpha \mid X\big)}{\partial b} = \frac{E\big(Y^2 \mid X, T=-1\big)}{c^2} \cdot \big[c - P\,(T=-1)\big].$$

Clearly, the sign of the derivatives depend on the value of c, and there are four possible cases.

Case 1

$$c > \max\left\{P\,(T=1), P\,(T=-1)\right\}$$

We have $\frac{\partial Var(\delta TY/c\alpha|X)}{\partial a} > 0, \frac{\partial Var(\delta TY/c\alpha|X)}{\partial b} > 0$. Hence, a and b should be as small as possible. Meanwhile, a and b have to be within a certain range such that δ is still a random variable, i.e., values in (2.5) are all between 0 and 1. Note that

$$P\big(\delta = a\alpha \mid T = 1\big) = \frac{P(T=1) - c}{(1-a)\,P(T=1)},$$

$$\frac{\partial P\big(\delta = a\alpha \mid T = 1\big)}{\partial a} = \frac{P(T=1) - c}{P(T=1)} \cdot \frac{1}{(1-a)^2} < 0.$$

So a achieves the smallest value when $P\big(\delta = a\alpha \mid T = 1\big) = 1$. Similarly, b achieves the smallest value when $P\big(\delta = b\alpha \mid T = -1\big) = 1$.

Thus, $Var\big(\delta TY/c\alpha \mid X\big)$ is minimized when

$$a = \frac{c}{P(T=1)} \text{ and } b = \frac{c}{P(T=-1)}.$$

Case 2 and Case 3

$$\min\left\{P(T=1), P(T=-1)\right\} < c < \max\left\{P(T=1), P(T=-1)\right\}$$

$$\text{or } c < \min\left\{P(T=1), P(T=-1)\right\}.$$

Following the same discussion as in *Case 1*, $Var\big(\delta TY/c\alpha \mid X\big)$ is minimized when

$$a = \frac{c}{P(T=1)} \text{ and } b = \frac{c}{P(T=-1)}.$$

Case 4

$$c = P(T = 1) \text{ or } c = P(T = -1).$$

If $c = P(T = 1)$, following the same discussion as in *Case 1*, we have $Var(\delta TY/c\alpha \mid X)$ is minimized when $b = \frac{c}{P(T=-1)}$, but a can be an arbitrary positive number (of course, 1 is not included by definition). Similarly, if $c = P(T = -1)$, we need $a = \frac{c}{P(T=1)}$, but b can be an arbitrary positive number except 1.

In all of the above four cases, the minimized variance for a given X is

$$\begin{aligned} Var(\delta TY/c\alpha \mid X) =& \frac{E(Y^2 \mid X, T = 1)}{P(T = 1)} + \frac{E(Y^2 \mid X, T = -1)}{P(T = -1)} \\ & - \left[E(Y \mid X, T = 1) - E(Y \mid X, T = -1)\right]^2. \end{aligned} \tag{2.10}$$

Clearly, the minimum variance does not depend on c, and we may let $c = 1$ for simplicity. Hence, a typical choice of a, b and c that minimize the variance of $\frac{\delta YT}{c\alpha}$ for any given X is

$$a = \frac{c}{P(T = 1)}, \; b = \frac{c}{P(T = -1)}, \text{ and } c = 1$$

Remark Whether the variance of $\frac{\delta YT}{c\alpha}$ for a given X is affected by the value of c depends on the value of a and b.

2.3.4 Special Choices of Parameter a, b and c

Just as the weighted least squares method may assign different weights to different observations, the number of replications of each observation in the working dataset may also be modified when needed. For example, we may want to have a balanced working dataset, meaning that the number of patients in the treatment arm is the same as that in the control arm. Or we may want to keep the proportion between the two arms in the working dataset the same as that of the original dataset. Meanwhile, we also want $Var\left(\delta TY/c\alpha \mid X\right)$ as small as possible, and δ still has to be a random variable.

2.3.4.1 Choice of Parameters for a Balanced Working Dataset

When the number of patients in the treatment arm and the control arm are the same in the working dataset, the following condition has to be satisfied

$$\alpha^2 P\big(T=1,\delta=\alpha\big)+a^2\alpha^2 P\big(T=1,\delta=a\alpha\big)=\alpha^2 P\big(T=-1,\delta=\alpha\big)+b^2\alpha^2 P\big(T=-1,\delta=b\alpha\big) \tag{2.11}$$

Based on the distribution of δ, Equation (2.11) leads to the following

$$\alpha^2\cdot\frac{c-aP\left(T=1\right)}{1-a}+a^2\alpha^2\cdot\frac{P\left(T=1\right)-c}{1-a}=\alpha^2\cdot\frac{c-bP\left(T=-1\right)}{1-b}+b^2\alpha^2\cdot\frac{P\left(T=-1\right)-c}{1-b}$$

$$\Longleftrightarrow a\big(c-P\left(T=1\right)\big)=b\big(c-P\left(T=-1\right)\big)$$

Hence, to have two balanced arms, we need

$$\frac{a}{b} = \frac{c - P(T=-1)}{c - P(T=1)}. \tag{2.12}$$

We need to minimize $Var\big(\delta TY/c\alpha \mid X\big)$ when (2.12) holds. From (2.12), let

$$a = k\big(c - P(T=-1)\big)\ ,\ b = k\big(c - P(T=1)\big) \tag{2.13}$$

where $k \neq 0$ is a constant. Substituting a, b in (2.9) with those in (2.13), the variance of $\frac{\delta YT}{c\alpha}$ for a given X becomes

$$\begin{aligned} Var\big(\delta TY/c\alpha \mid X\big) \;=\; & \frac{c + kc^2 - kc + kP(T=-1)\,P(T=1)}{c^2} \cdot \Big[E\left(Y^2 \mid X, T=1\right) \\ & + E\left(Y^2 \mid X, T=-1\right)\Big] - \big[E\big(Y \mid X, T=1\big) - E\big(Y \mid X, T=-1\big)\big]^2. \end{aligned}$$

And its derivative with respect to k is

$$\frac{\partial Var\big(\delta TY/c\alpha \mid X\big)}{\partial k} = \frac{E\big(Y^2 \mid X, T=1\big) + E\big(Y^2 \mid X, T=-1\big)}{c^2} \cdot \big[c - P(T=1)\big] \cdot \big[c - P(T=-1)\big]. \tag{2.14}$$

The sign of (2.14) depends on c. Note that a, $b > 0$ and c cannot be $P(T=1)$ or $P(T=-1)$ according to (2.12). Thus, there are two possible cases.

Case 1

$$c > \max\left\{P(T=1), P(T=-1)\right\}$$

We have $\frac{\partial Var\left(\delta TY/c\alpha|X\right)}{\partial k} > 0$, so k should be as small as possible. Meanwhile, a and b have to be within a certain range such that δ is still a random variable. Note that

$$P\big(\delta = a\alpha \mid T = 1\big) = \frac{P\left(T = 1\right) - c}{\big[1 + k\left(P\left(T = -1\right) - c\right)\big]P\left(T = 1\right)},$$

$$\frac{\partial P\big(\delta = a\alpha \mid T = 1\big)}{\partial k} = \frac{P\left(T = 1\right) - c}{P\left(T = 1\right)} \cdot \big[1 + k\left(P\left(T = -1\right) - c\right)\big]^{-2} \cdot \big[c - P\left(T = -1\right)\big] < 0.$$

Therefore, $P\big(\delta = a\alpha \mid T = 1\big) \leq 1 \Rightarrow k \geq \frac{c}{cP\left(T = 1\right) - P\left(T = 1\right)P\left(T = -1\right)}$

Similarly, we have

$$P\big(\delta = b\alpha \mid T = -1\big) = \frac{P\left(T = -1\right) - c}{\big[1 + k\left(P\left(T = 1\right) - c\right)\big]P\left(T = -1\right)},$$

$$\frac{\partial P\big(\delta = b\alpha \mid T = -1\big)}{\partial k} = \frac{P\left(T = -1\right) - c}{P\left(T = -1\right)} \cdot \big[1 + k\left(P\left(T = 1\right) - c\right)\big]^{-2} \cdot \big[c - P\left(T = 1\right)\big] < 0.$$

Therefore, $P\big(\delta = b\alpha \mid T = -1\big) \leq 1 \Rightarrow k \geq \frac{c}{cP\left(T = -1\right) - P\left(T = 1\right)P\left(T = -1\right)}$.

Hence, to minimize $Var\big(\delta TY/c\alpha \mid X\big)$, we need

$$k = \begin{cases} \dfrac{c}{cP\left(T = 1\right) - P\left(T = 1\right)P\left(T = -1\right)} & \text{if } \ P\left(T = 1\right) < P\left(T = -1\right) \\ \dfrac{c}{cP\left(T = -1\right) - P\left(T = 1\right)P\left(T = -1\right)} & \text{if } \ P\left(T = 1\right) > P\left(T = -1\right) \end{cases}$$

Case 2

$$c < \min\left\{P\left(T = 1\right), P\left(T = -1\right)\right\}$$

Following the same discussion as in *Case 1*, $\mathrm{Var}\left(\delta TY/c\alpha \mid X\right)$ is minimized when

$$k = \begin{cases} \dfrac{c}{cP\left(T=1\right) - P\left(T=1\right)P\left(T=-1\right)} & \text{if } \ P\left(T=1\right) < P\left(T=-1\right) \\ \dfrac{c}{cP\left(T=-1\right) - P\left(T=1\right)P\left(T=-1\right)} & \text{if } \ P\left(T=1\right) > P\left(T=-1\right) \end{cases}$$

In both *Case 1* and *Case 2*, the minimized variance for a given X after simplification is

$$\min \ \mathrm{Var}\left(\frac{\delta TY}{c\alpha} \mid X\right) = \begin{cases} \dfrac{1}{P\left(T=1\right)} \cdot \left[E\left(Y^2 \mid T=1, X\right) + E\left(Y^2 \mid T=-1, X\right)\right] \\ -\left[E\left(Y \mid T=1, X\right) - E\left(Y \mid T=-1, X\right)\right]^2, \\ \qquad\qquad when \ \ P\left(T=1\right) < P\left(T=-1\right) \\ \dfrac{1}{P\left(T=-1\right)} \cdot \left[E\left(Y^2 \mid T=1, X\right) + E\left(Y^2 \mid T=-1, X\right)\right] \\ -\left[E\left(Y \mid T=1, X\right) - E\left(Y \mid T=-1, X\right)\right]^2, \\ \qquad\qquad when \ \ P\left(T=1\right) > P\left(T=-1\right). \end{cases} \tag{2.15}$$

The minimized variance in (2.15) are not dependent on c. So, we may let $c = 1$ for simplicity. Comparing (2.15) with (2.10), the two minimized variances have very similar form, and the actual difference between the two depends on how unbalanced the original dataset is. Based on

the above discussion, a typical choice of a, b and c that minimize the variance of $\frac{\delta TY}{c\alpha}$ for any given X while providing a balanced working dataset is

$$a = kP(T=1)\ ,\ b = kP(T=-1)\ ,\ c = 1$$

$$\text{where } k = \begin{cases} \dfrac{1}{P^2(T=1)}, & \text{if } P(T=1) < P(T=-1) \\ \dfrac{1}{P^2(T=-1)}, & \text{if } P(T=1) > P(T=-1) \end{cases}$$

2.3.4.2 Choice of Parameters for a Proportional Working Dataset

When the proportion between the two arms in the working dataset is the same as that of the original dataset, the following condition has to be satisfied

$$\frac{\alpha^2 P(T=1)\,P\big(\delta=\alpha \mid T=1\big) + a^2\alpha^2 P(T=1)\,P\big(\delta=a\alpha \mid T=1\big)}{\alpha^2 P(T=-1)\,P\big(\delta=\alpha \mid T=-1\big) + b^2\alpha^2 P(T=-1)\,P\big(\delta=b\alpha \mid T=-1\big)} = \frac{P(T=1)}{P(T=-1)}. \quad (2.16)$$

Based on the distribution of δ, Equation (2.16) leads to

$$\frac{c + ca - aP(T=1)}{c + cb - bP(T=-1)} = \frac{P(T=1)}{P(T=-1)} \quad (2.17)$$

We need to minimize $Var\big(\frac{\delta TY}{c\alpha} \mid X\big)$ when (2.17) holds. Suppose

$$c + ca - aP(T=1) = kP(T=1), c + cb - bP(T=-1) = kP(T=-1) \text{ for some k}$$

Then we have

$$a = \frac{kP(T=1)-c}{c-P(T=1)}, b = \frac{kP(T=-1)-c}{c-P(T=-1)}$$

The variance of $\frac{\delta TY}{c\alpha}$ for a given X becomes

$$\begin{aligned} Var\left(\frac{\delta TY}{c\alpha} \mid X\right) &= \frac{kP(T=1)}{c^2}E\left(Y^2 \mid T=1,X\right) + \frac{kP(T=-1)}{c^2}E\left(Y^2 \mid T=-1,X\right) \\ &\quad -\left[E(Y \mid T=1,X) - E(Y \mid T=-1,X)\right]^2 \end{aligned}$$

and its derivative with respect to k is

$$\frac{\partial Var\left(\frac{\delta TY}{c\alpha} \mid X\right)}{\partial k} = \frac{1}{c^2}E\left(Y^2 \mid X\right) > 0.$$

Thus, k should be as small as possible. Note that δ has to be a random variable.

$$P\left(\delta = a\alpha \mid T=1\right) = \frac{\left[P(T=1)-c\right]^2}{\left[kP(T=1)+P(T=1)-2c\right]\cdot P(T=1)}$$

$$P\left(\delta = b\alpha \mid T=-1\right) = \frac{\left[P(T=-1)-c\right]^2}{\left[kP(T=-1)+P(T=-1)-2c\right]\cdot P(T=-1)}$$

$$\frac{\partial P\left(\delta = a\alpha \mid T=1\right)}{\partial k} < 0$$

$$\frac{\partial P\left(\delta = b\alpha \mid T=-1\right)}{\partial k} < 0$$

Hence, to minimize $Var\left(\frac{\delta TY}{c\alpha} \mid X\right)$, we need

$$k = \max\left\{\frac{c^2}{\left[P\left(T=1\right)\right]^2}, \frac{c^2}{\left[P\left(T=-1\right)\right]^2}\right\}$$

The minimized variance is thus

$$\min\ Var\left(\frac{\delta TY}{c\alpha} \mid X\right) = \begin{cases} \dfrac{1}{P\left(T=1\right)}E\left(Y^2 \mid X, T=1\right) + \dfrac{P\left(T=-1\right)}{\left[P\left(T=1\right)\right]^2}E\left(Y^2 \mid X, T=-1\right) \\ -\left[E\left(Y \mid X, T=1\right) - E\left(Y \mid X, T=-1\right)\right]^2, \\ \qquad\qquad when\ \ P\left(T=1\right) < P\left(T=-1\right) \\ \\ \dfrac{P\left(T=1\right)}{\left[P\left(T=-1\right)\right]^2}E\left(Y^2 \mid X, T=1\right) + \dfrac{1}{P\left(T=-1\right)}E\left(Y^2 \mid X, T=-1\right) \\ -\left[E\left(Y \mid X, T=1\right) - E\left(Y \mid X, T=-1\right)\right]^2, \\ \qquad\qquad when\ \ P\left(T=1\right) > P\left(T=-1\right) \end{cases} \tag{2.18}$$

The minimized variance does not depend on c, and we let $c = 1$ for simplicity. Note that (2.18) also has a similar form as (2.10), and the actual difference depends on the proportion of the two arms in the original dataset. Thus, a typical choice of a, b and c that minimize the variance of $\frac{\delta TY}{c\alpha}$ for any given X while keeping the proportion between the two arms is

$$a = \frac{kP\left(T=1\right)-1}{P\left(T=-1\right)}\ ,\ b = \frac{kP\left(T=-1\right)-1}{P\left(T=1\right)},\ c = 1,$$
$$\text{where } k = \max\left\{\frac{1}{P^2\left(T=1\right)}, \frac{1}{P^2\left(T=-1\right)}\right\}$$

2.4 Interaction Filtering for the Survival Response

2.4.1 Extension to the Survival Response

One advantage of constructing a working model and a working dataset in the proposed way is that it can be intuitively extended to other responses. Here, we extend the interaction filtering procedure to the survival response.

i. For patient i from 1 to N, replicate the corresponding observation δ_i^2 times, where δ_i is generated from the distribution of δ. The generated new dataset is called the working dataset.

ii. Use the working dataset from step i and the working model below to perform variable selection with an appropriate method such as LASSO under the context of the Cox proportional hazard model (32). The selected variables are then the ones that potentially interact with the treatment.

$$\log\frac{h(t \mid X_i)}{h_0(t)} = 2c\alpha\gamma' X_i \cdot T_i/\delta_i \tag{2.19}$$

Compared with the case of the continuous response, the survival response case has exactly the same working dataset generating procedure, but the working model is now based on the Cox regression. Here, we use the same choice of parameter a, b and c as in the continuous case for simplicity.

2.4.2 Justification of the Objective Function

Though the extension is intuitive, we need to justify that the objective function based on the working dataset and the working model is reasonable for filtering the interactions. The proposed procedure is based on the Cox model, the objective function is thus the partial likelihood, which is as follows when using the working model and the working dataset

$$\max_{f} \sum_{i=1}^{N} \delta_i^2 \xi_i \left[c\alpha f(X_i) T_i/\delta_i - \log \left(\sum_{j=1}^{N} \delta_j^2 \cdot \exp\big(c\alpha f(X_j) T_j/\delta_j\big) \cdot I\big(Y_j \geq Y_i\big) \right) \right] \tag{2.20}$$

$$\text{subject to } f_i \in F_i = \{2\gamma' X_i \mid r \in R^p\}$$

where Y is the observed time, and ξ is the event indicator with 1 indicating the event and 0 indicating the censoring. Also assume that ξ is independent of T, X and $\tilde{Y}$, where $\tilde{Y}$ is the survival time, and equals Y when the observation is not censored.

The objective function (2.20) can be rewritten as follows

$$\begin{aligned}
& \max_{f} \frac{1}{N} \sum_{i=1}^{N} \left[\xi_i c\alpha f(X_i) T_i \delta_i - \delta_i^2 \xi_i \cdot \log \left(\sum_{j=1}^{N} \delta_j^2 \cdot \exp\big(c\alpha f(X_j) T_j/\delta_j\big) \cdot I\big(Y_j \geq Y_i\big) \right) \right] \\
\Longleftrightarrow \quad & \max_{f} \frac{1}{N} \sum_{i=1}^{N} \left[\xi_i c\alpha f(X_i) T_i \delta_i - \xi_i \delta_i^2 \cdot \log \left(\frac{1}{N} \sum_{j=1}^{N} \delta_j^2 \cdot \exp\big(c\alpha f(X_j) T_j/\delta_j\big) \cdot I\big(\tilde{Y}_j \geq \tilde{Y}_i\big) \right) \right] \\
& \textit{where } \tilde{Y}_i \textit{ is the actual survival time} \\
\Longleftrightarrow \quad & \max_{f} \frac{1}{N} \sum_{i=1}^{N} \Bigg\{ \xi_i c\alpha f(X_i) T_i \delta_i \\
& \qquad - \int_0^{\infty} \xi_i \delta_i^2 \cdot \left[\log \left(\frac{1}{N} \sum_{j=1}^{N} \delta_j^2 \cdot \exp\big(c\alpha f(X_j) T_j/\delta_j\big) \cdot I\big(\tilde{Y}_j \geq t\big) \right) \right] \Delta\left(t - \tilde{Y}_i\right) dt \Bigg\}
\end{aligned}$$

Thus, the objective function has the following general form

$$\max_{f} \left\{ E\Big(\xi c\alpha f(X)\, T\delta\Big) - E\left[\int_0^{\infty} \xi\delta^2 \log\left[E\left(\delta^2 \cdot exp\big(c\alpha f(X)\, T/\delta\big) \cdot I\left(\tilde{Y} \geq t\right)\right)\right] \cdot \Delta\left(t - \tilde{Y}\right) dt\right]\right\} \tag{2.21}$$

Note that ξ is independent of T, hence,

$$E\big[\xi c\alpha f(X)\, T\delta\big] = E_X\big[c\alpha f(X) \cdot E(\delta T \mid X) \cdot E\big(\xi \mid X\big)\big] = 0$$

Thus, the objective function becomes

$$\min_{f} l^*(f) = \left\{ \int_0^{\infty} E\left[\xi\delta^2 \Delta\left(t - \tilde{Y}\right)\right] \log E\left[\delta^2 exp\left(c\alpha f(X)\, T/\delta\right) I\left(\tilde{Y} \geq t\right)\right] dt \right\} \tag{2.22}$$

$$\text{subject to } f \in \mathcal{F} = \{2\gamma' X \mid \gamma \in R^p\}$$

Let $f_0(X)$ and $f^*(X)$ be the respective minimizer of (2.22) when no specific form of $f(X)$ is assumed and when $f(X)$ is in the space of $\mathcal{F}$. We can show that there is an association between the individual treatment effect and $f_0(X)$ regardless of the true underlying model. Hence, considering that $f^*(X)$ is a surrogate of $f_0(X)$ in a functional space, it is reasonable to use $f^*(X)$ for interaction filtering.

To find the minimum of $l^*(f)$, let

$$\frac{\partial l^*\big(f(X) + \epsilon\eta(X)\big)}{\partial \epsilon} \Big|_{\epsilon=0} = 0$$

where $\eta(X)$ is an arbitrary function of X. Thus, $f_0(X)$ satisfies the following condition

$$E_X\left\{\int_0^\infty E\left[\xi\delta^2\Delta\left(t-\tilde{Y}\right)\right]\cdot\frac{E\left[\delta\exp\left(c\alpha f_0(X)T/\delta\right)I\left(\tilde{Y}\geq t\right)c\alpha T\mid X\right]}{E\left[\delta^2\exp\left(c\alpha f_0(X)T/\delta\right)I\left(\tilde{Y}\geq t\right)\right]}dt\cdot\eta(X)\right\}=0,\forall\eta(X)$$

which is equivalent to

$$\int_0^\infty E\left[\xi\delta^2\Delta\left(t-\tilde{Y}\right)\right]\cdot\frac{E\left[\delta\exp\left(c\alpha f_0(X)T/\delta\right)I\left(\tilde{Y}\geq t\right)T\mid X\right]}{E\left[\delta^2\exp\left(c\alpha f_0(X)T/\delta\right)I\left(\tilde{Y}\geq t\right)\right]}dt\cdot=0. \tag{2.23}$$

Note that

$$\begin{aligned}
& E\left[\delta\exp\left(c\alpha f_0(X)T/\delta\right)I\left(\tilde{Y}\geq t\right)T\mid X\right] \\
& \delta\perp\!\!\!\perp\tilde{Y}\mid T \\
= \; & E\left[\delta\exp\left(c\alpha f_0(X)/\delta\right)\mid X,T=1\right]\cdot E\left[I\left(\tilde{Y}\geq t\right)\mid X,T=1\right]\cdot P(T=1) \\
& -E\left[\delta\exp\left(-c\alpha f_0(X)/\delta\right)\mid X,T=-1\right]\cdot E\left[I\left(\tilde{Y}\geq t\right)\mid X,T=-1\right]\cdot P(T=-1) \\
= \; & \phi\left(a,f_0(X)\right)\cdot S(t\mid X,T=1)-\phi\left(b,-f_0(X)\right)\cdot S(t\mid X,T=-1)
\end{aligned}$$

$$\begin{aligned}
\text{where } \phi\left(u,f_0(X)\right)=\alpha\cdot\exp\left(cf_0(X)\right)\cdot\frac{c-u\cdot\left[I(u=a)P(T=1)+I(u=b)P(T=-1)\right]}{1-u} \\
+u\cdot\alpha\cdot\exp\left(cf_0(X)/u\right)\cdot\frac{\left[I(u=a)P(T=1)+I(u=b)P(T=-1)\right]-c}{1-u}
\end{aligned}$$

Also Note that

$$\begin{aligned}
& E\left[\xi\delta^2\Delta\left(t-\tilde{Y}\right)\right] \qquad\qquad (\xi \perp\!\!\!\perp T, \tilde{Y}) \\
= \; & E(\xi)\cdot\left\{E\left[\delta^2 \mid T=1\right]\cdot P(T=1)\cdot E\left[\Delta\left(t-\tilde{Y}\right) \mid T=1\right]\right. \\
& \qquad \left. + E\left[\delta^2 \mid T=-1\right]\cdot P(T=-1)\cdot E\left[\Delta\left(t-\tilde{Y}\right) \mid T=-1\right]\right\} \\
= \; & E(\xi)\cdot E\left(\delta^2\right)\cdot\left\{P^*(T=1)\cdot E\left[\Delta\left(t-\tilde{Y}\right) \mid T=1\right] + P^*(T=-1)\cdot E\left[\Delta\left(t-\tilde{Y}\right) \mid T=-1\right]\right\}
\end{aligned} \tag{2.24}$$

where

$$P^*(T=1) = E\left[\delta^2 \mid T=1\right]\cdot P(T=1)/E\left(\delta^2\right)$$

$$P^*(T=-1) = E\left[\delta^2 \mid T=-1\right]\cdot P(T=-1)/E\left(\delta^2\right)$$

and $P^*(T=1)$ and $P^*(T=-1)$ are the respective proportion of the treatment arm and the control arm in the working dataset. Hence, (2.24) can be further written as

$$E(\xi)\cdot E\left(\delta^2\right)\cdot g^*_{\tilde{Y}}(t)$$

where $g^*_{\tilde{Y}}(t)$ is the marginal density of $\tilde{Y}$ in the working data.

Hence, Condition (2.23) is equivalent to the following

$$\int_0^\infty g^*_{\tilde{Y}}(t)\,\frac{\phi\left(a, f_0(X)\right)\cdot S\left(t \mid X, T=1\right)}{G\left(f_0(X), t\right)}dt = \int_0^\infty g^*_{\tilde{Y}}(t)\,\frac{\phi\left(b, -f_0(X)\right)\cdot S\left(t \mid X, T=-1\right)}{G\left(f_0(X), t\right)}dt$$

where $G\left(f_0(X), t\right) = E\left[\delta^2 exp\left(c\alpha f_0(X) T/\delta\right) I\left(\tilde{Y} \geq t\right)\right]$.

$$\Phi\left(f_0(X)\right) = \frac{E^*_{\tilde{Y}}\left[\frac{S\left(\tilde{Y}|X,T=-1\right)}{G\left(f_0(X),\tilde{Y}\right)}\right]}{E^*_{\tilde{Y}}\left[\frac{S\left(\tilde{Y}|X,T=1\right)}{G\left(f_0(X),\tilde{Y}\right)}\right]} \tag{2.25}$$

where $\Phi = \frac{\phi(a,f_0(X))}{\phi(b,-f_0(X))}$ is a monotone increasing function of $f_0(X)$.

The right hand side of Equation (2.25) can be considered as the representative of the individual treatment effect, thus (2.25) builds an association between the individual treatment effect and $f_0(X)$. Since $f^*(X)$ is a surrogate of $f_0(X)$ in a functional space, it is reasonable to use $f^*(X)$ for interaction filtering.

2.5 Remarks for Implementation

In practice, after building the working dataset and the working model, we use LASSO as the variable selection method. Other methods such as elastic net (33) can also be used. There are several small tricks that can further enhance the performance.

The minimum variance of $\frac{\delta T Y}{c\alpha}$ for a given X under the three different cases we propose for the continuous response are shown in (2.10), (2.15) and (2.18). It is clear that if the response is centered around 0 for every given X, then the value of the minimum variance can be reduced though the formulas remain unchanged. One way to make the response mean for every given X close to 0 is to subtract the estimated main effects from the response. Note that the main effects should be estimated when the treatment and the control arm are coded as 1 and -1 respectively (the estimation of the main effects are affected by the treatment code). This can be done by fitting a rough complete model like (2.1) with LASSO and estimating important

main effects. The estimation does not have to be accurate. As long as the response is more centered around 0 after reducing the main effects, the value of the minimum variance will be smaller. Actually, even simply subtracting the sample mean of the response will help. The idea of reducing the main effects is also intuitively correct for both the continuous and the survival response, since the working model will thus be closer to the true model and the smaller main effects after reduction are easier to cancel out.

We have a parameter α when building the working model and the working dataset, which helps make the number of replications of each observation an integer. This is only useful for illustration purpose such that the user will not get confused by replicating an observation a fraction of a time. When actually implementing the procedure, replicating an observation is equivalent to assigning a corresponding weight, so we do not have to make it an integer. Thus, we simply let $\alpha = 1$ in practice.

The LASSO method relies on a cross validation procedure to select the best tuning parameter. The cross validation procedure involves splitting the dataset randomly into several parts, so the selected best tuning parameter may be different if the dataset is split in a different way. In practice, we run the cross validation multiple times and use the median of the selected tuning parameters.

2.6 Simulation Studies

2.6.1 Simulation Setup

In this section, we perform a series of simulation studies to investigate the performance of the proposed method under various conditions. The following two methods are compared

(i) The proposed interaction filtering procedure based on the working model and the working dataset, with LASSO as the variable selection method (IFL)

(ii) The ordinary multivariate regression procedure based on the complete model (2.1) with LASSO as the variable selection method (OL)

The following metrics are used for comparing the performance. We borrow the name Sensitivity and Positive Predictive Value, though the situation here is a little different from where the terms are normally used. We call a covariate a true interaction covariate if there is a corresponding interaction term specified in the simulation model.

(i) Sensitivity

Definition The probability that a covariate is selected as an interaction covariate given that it is a true interaction covariate.

Calculation

$$\frac{\left|\left\{X_i : X_i \text{ is a true interacation covariate \& } X_i \text{ is selected as an interaction covariate}\right\}\right|}{\left|\left\{X_i : X_i \text{ is a true interacation covariate }\right\}\right|}$$

(ii) Positive Predictive Value (PPV)

Definition The probability that a selected interaction covariate is a true interaction covariate.

Calculation

$$\frac{\left|\left\{X_i : X_i \text{ is a true interacation covariate \& } X_i \text{ is selected as an interaction covariate}\right\}\right|}{\left|\left\{X_i : X_i \text{ is selected as an interaction covariate}\right\}\right|}$$

(iii) Type I Error

Definition The probability that at least one covariate is selected given that there is no true interaction covariate.

Calculation Running the simulation multiple times under the condition that there is no true interaction covariate, and calculate the proportion of times that at least one interaction covariate is selected.

(Note that the main effects selected by the OL method is not involved in the calculations.)

We use a similar data generating procedure as in Tian's paper (30). For the continuous response, the following model is used to generate the dataset

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon \tag{2.26}$$

where p is set to be 100, and the covariates $(X_1, X_2, ..., X_{100})$ follow a multivariate normal distribution with a compound symmetric variance-covariance matrix

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{100} \end{pmatrix} \sim N\left(\begin{pmatrix} 5 \\ 5 \\ \vdots \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \right)$$

The means of the covariates are set to 5 to mimic the real world data which usually have positive measurements. The random error ϵ follows Normal (0, 1). The probability that a patients belongs to the treatment arm is $2/3$, i.e., $P(T = 1) = 2/3$, $P(T = -1) = 1/3$. When there exist true interaction covariates, the values of $\gamma's$ are as follows

$$\gamma_1 = 1, \gamma_2 = -1, \gamma_3 = 2, \gamma_4 = -2, \text{ and all other } \gamma = 0$$

Thus, the four true interaction covariates are X_1, X_2, X_3 and X_4.

For the survival response, we generate the data from a modification of (2.26)

$$Y = exp\left\{ 0.1 \cdot \left(\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon \right) \right\} \tag{2.27}$$

All the settings and parameters in the survival response case are exactly the same as in the continuous response case. The event indicator is randomly generated to induce 20% censoring rate.

We consider the following different scenarios for simulation.

(i) $\beta_5 = 1$, $\beta_6 = -1$, $\beta_7 = 2$, $\beta_8 = -2$, all other $\beta = 0$

$p = 0.2$, sample size $N = 50, 100, 200$

(ii) $\beta_5 = 1$, $\beta_6 = -1$, $\beta_7 = 2$, $\beta_8 = -2$, all other $\beta = 0$

$p = 0.5$, sample size $N = 50, 100, 200$

(iii) $\beta_5 = 2$, $\beta_6 = -2$, $\beta_7 = 4$, $\beta_8 = -4$, all other $\beta = 0$

$p = 0.2$, sample size $N = 50, 100, 200$

(iv) $\beta_5 = 2$, $\beta_6 = -2$, $\beta_7 = 4$, $\beta_8 = -4$, all other $\beta = 0$

$p = 0.5$, sample size $N = 50, 100, 200$

Scenario (i) and (ii) represent the cases that the main effects are moderate, while Scenario (iii) and (iv) represent the cases that the main effects are relatively large.

2.6.2 Simulation Result for the Continuous Response

2.6.2.1 Sensitivity and Positive Predictive Value

We conductded 500 runs of simulation for every case mentioned above. In each run, a dataset is generated, and both methods (IFL and OL) are applied to select interaction covariates. The corresponding Sensitivity and Positive Predictive Value are calculated. After having results from all the 500 runs, boxplots are generated to show the average performance.

Figure 1 and Figure 2 are the boxplots of Sensitivity and Positive Predictive Value when the main effects are moderate (Scenario (i) and (ii)). We can see that the two methods perform

almost the same in Sensitivity. The sample size has a positive impact on the Sensitivity, while the correlation among covariates has a negative impact. However, when it comes to the Positive Predictive Value, our method has a better performance, and this advantage becomes bigger as the sample size increases.

Figure 3 and Figure 4 are the boxplots of Sensitivity and Positive Predictive Value when the main effects are relatively large (Scenario (iii) and (iv)). Overall, we can see the same performance pattern as when the main effects are moderate. Comparing Figure 3 and Figure 1, our method essentially maintains the same performance, while the OL method surprisingly has a slightly improved Sensitivity when the main effects increase. A possible reason is that the OL method has to handle both the main effects and the interaction effects, and larger main effects can help it better select and estimate those main effects, which in turn helps select interaction covariates. As for the Positive Predictive Value, our methods still perform better, especially when the sample size becomes larger.

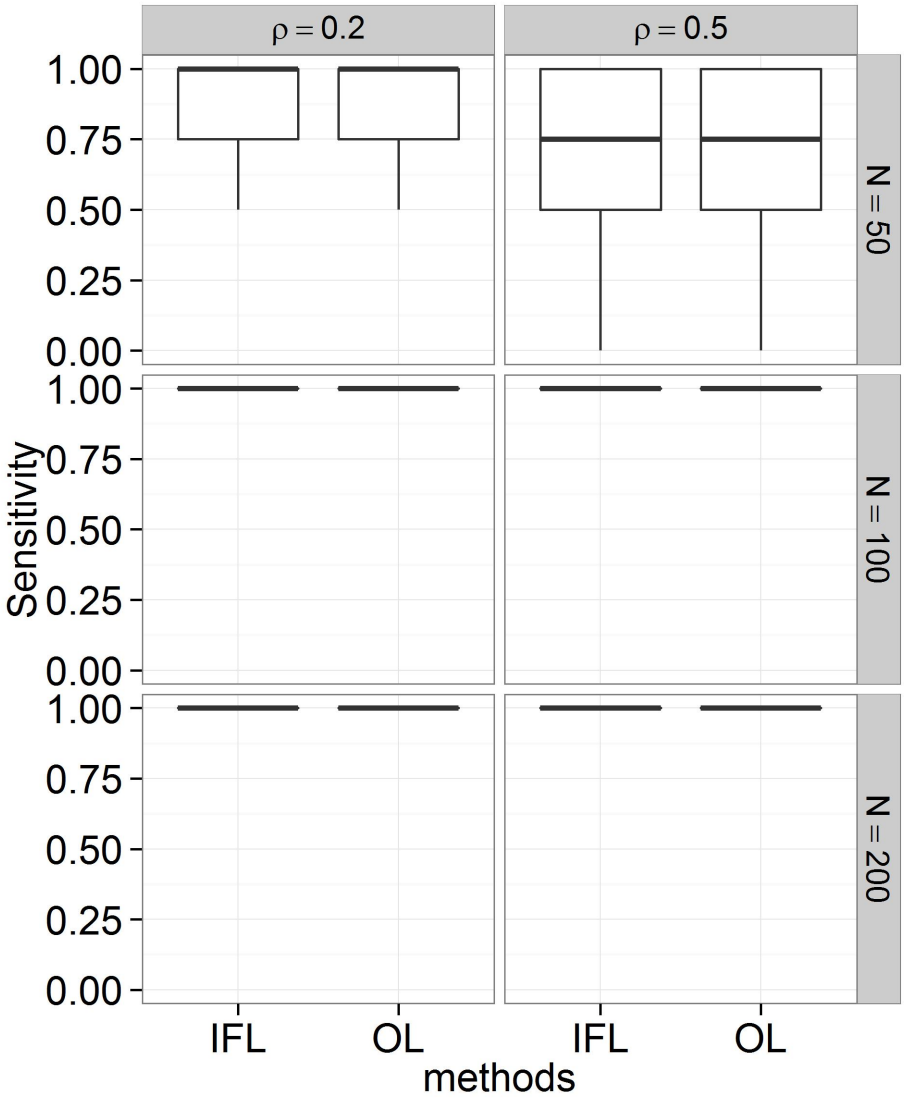

Figure 1. Sensitivity when the main effects are moderate (Scenario (i) and (ii), Continuous).

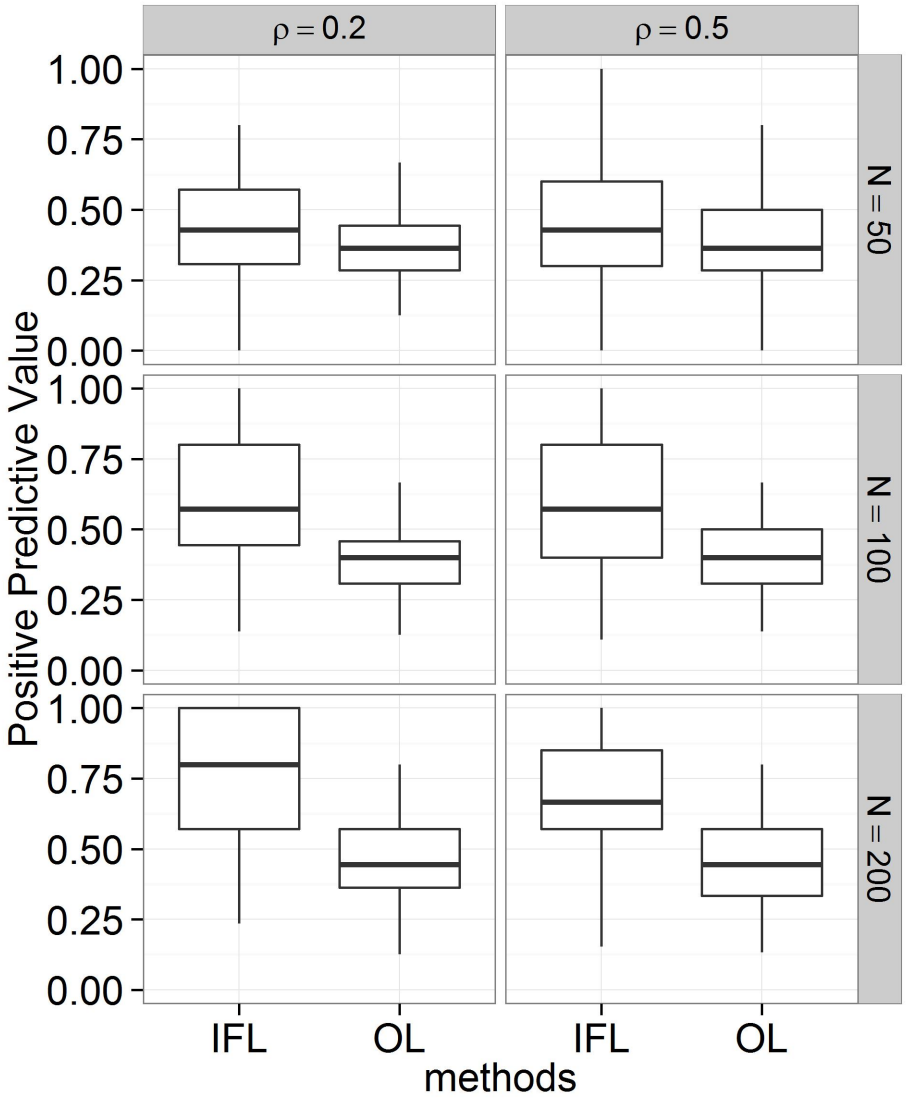

Figure 2. PPV when the main effects are moderate (Scenario (i) and (ii), Continuous).

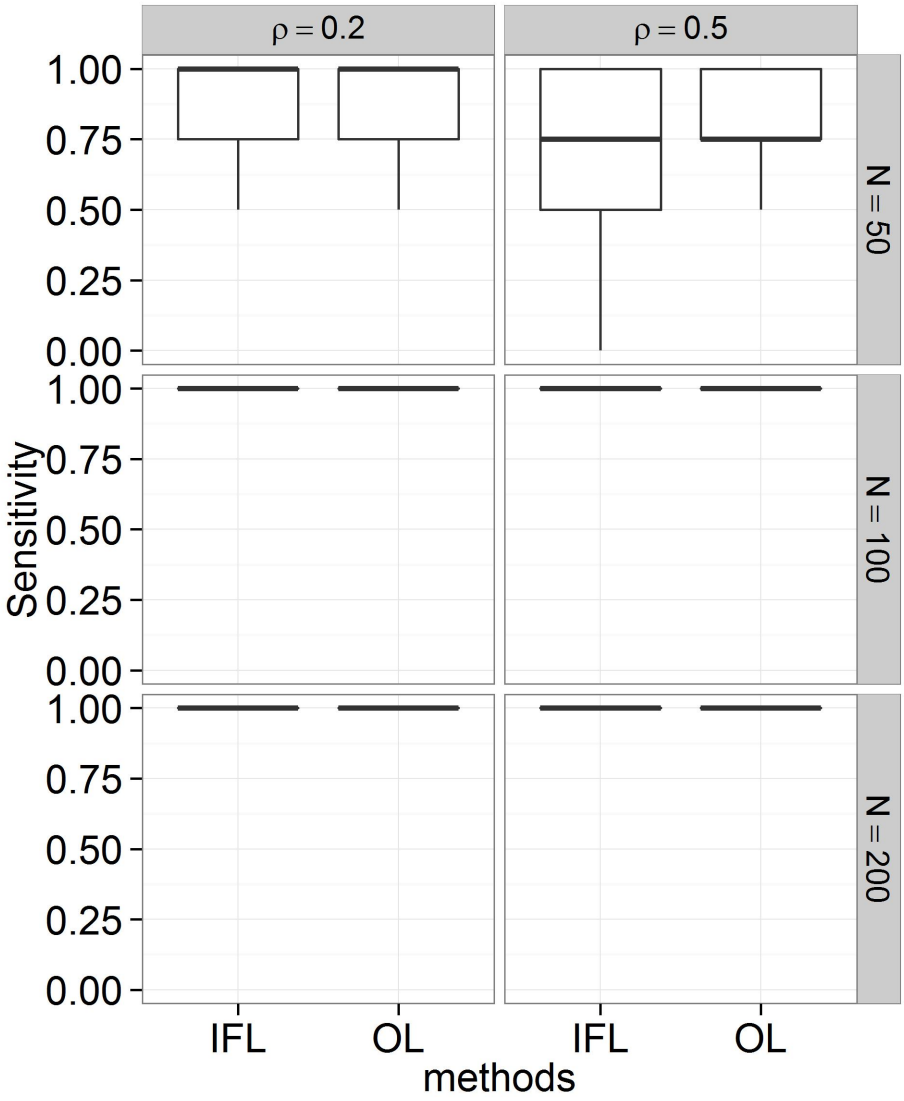

Figure 3. Sensitivity when the main effects are large (Scenario (iii) and (iv), Continuous).

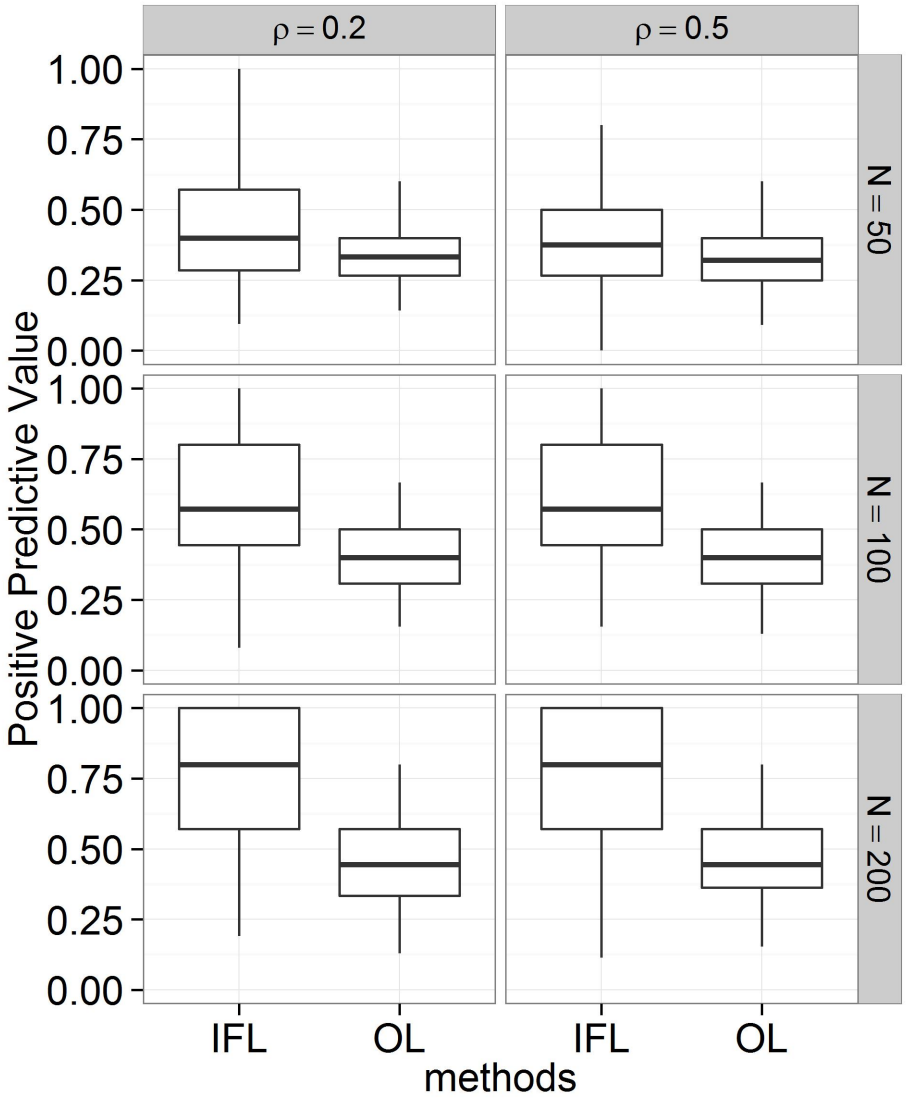

Figure 4. PPV when the main effects are large (Scenario (iii) and (iv), Continuous).

2.6.2.2 Type I Error

In the pharmaceutical area, a false discovery could waste a lot of time and resources if a separate trial is conducted to verify the result. Thus, it is interesting to find out how the methods perform when there are actually no interaction effects.

The same model and data generating procedure are used except that the interaction coefficients $\gamma's$ are all set to 0. We conducted 500 runs of simulation for every case. In each run, a dataset is generated, and both methods (IFL and OL) are applied to select interaction covariates. If a method reports that some interaction covariates are selected, a so-called error event is recorded for that method. The proportion of error events out of 500 runs is calculated as the estimated Type I Error for each method.

Since whether the main effects are small or relatively large does not make much difference in performance, we thus only show the results for the cases with large main effects here to avoid repetition. According to Table I, our method is significantly better than the OL method in all cases. The OL method has Type I Errors almost close to 1, and improves slowly when the sample size increases. To the contrary, our method has a very small Type I Error, and improves significantly when the sample size increases.

TABLE I

TYPE I ERROR (LARGE MAIN EFFECTS, CONTINUOUS)

p	N	Method	Type I Error
0.2	50	IFL	0.290
		OL	0.976
	100	IFL	0.052
		OL	0.968
	200	IFL	0.002
		OL	0.906
0.5	50	IFL	0.280
		OL	0.984
	100	IFL	0.040
		OL	0.974
	200	IFL	0.002
		OL	0.934

2.6.3 Simulation Result for the Survival Response

2.6.3.1 Sensitivity and Positive Predictive Value

Similar to the continuous response case, we conducted 500 runs of simulation for every case mentioned above. In each run, a dataset is generated, and both methods (IFL and OL) are applied to select interaction covariates. The corresponding Sensitivity and Positive Predictive Value are calculated. After having results from all the 500 runs, boxplots are generated to show the average performance.

Figure 5 and Figure 6 show the results when the main effects are moderate. When the sample size is very small ($N = 50$), the OL method beats our method in both Sensitivity and Positivity Predictive Value. Note that around half of 500 runs of simulation for each method cannot select any interaction covariates, so the Positive Predictive Value of these runs cannot be calculated and has to be excluded from the boxplot. When the sample size increases, our methods perform better in both Sensitivity and Positivity Predictive Value.

Figure 7 and Figure 8 show the results when the main effects are relatively large. The OL method beats our method in Sensitivity when the sample size is very small ($N = 50$), but our method catches up and outperforms the OL method as the sample size increases. As for Positive Predictive Value, our method has a better performance especially when the sample size is large.

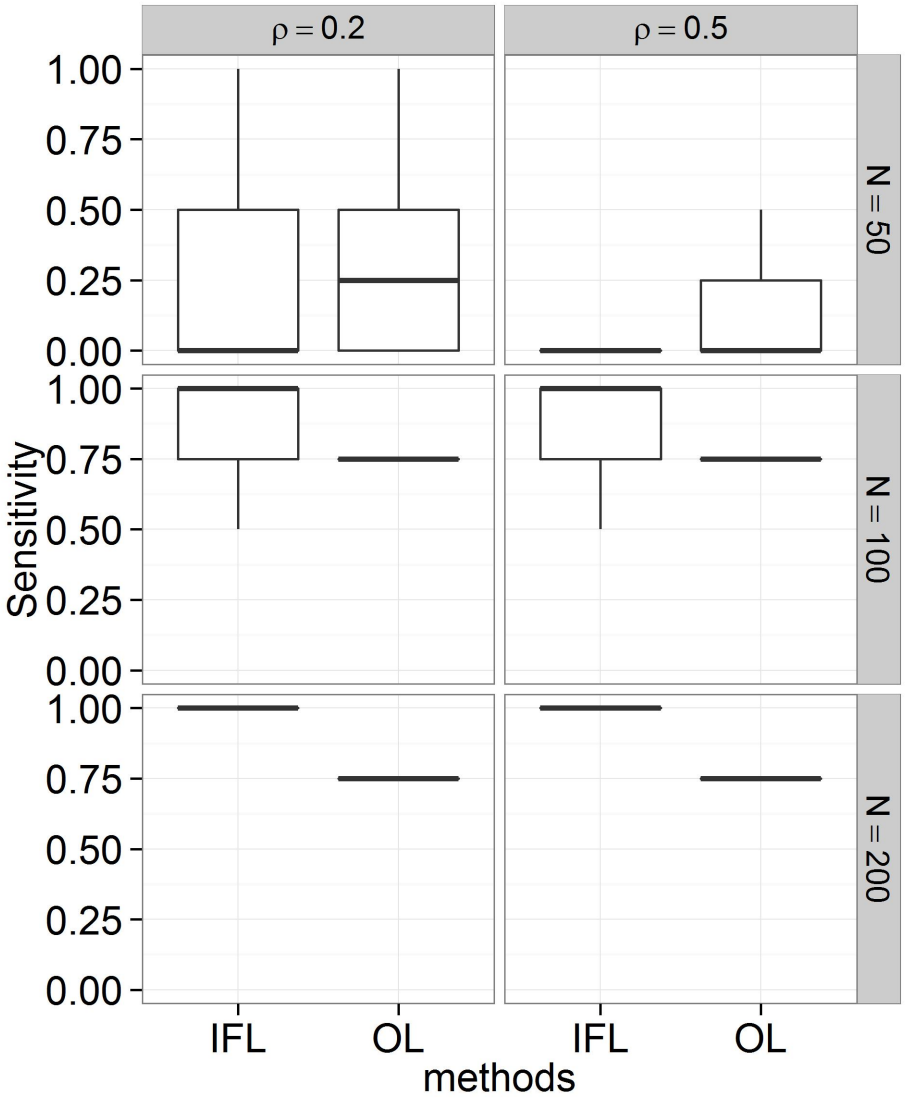

Figure 5. Sensitivity when the main effects are moderate (Scenario (i) and (ii), Survival).

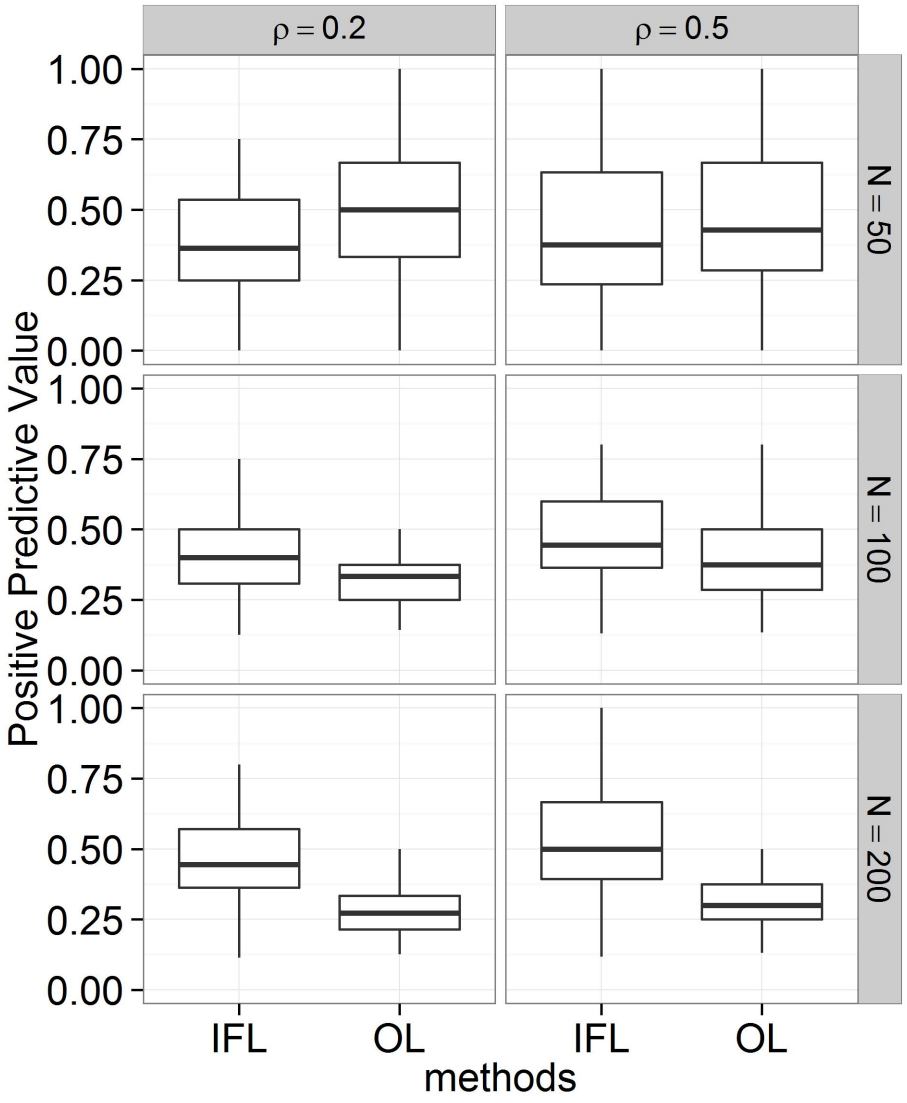

Figure 6. PPV when the main effects are moderate (Scenario (i) and (ii), Survival).

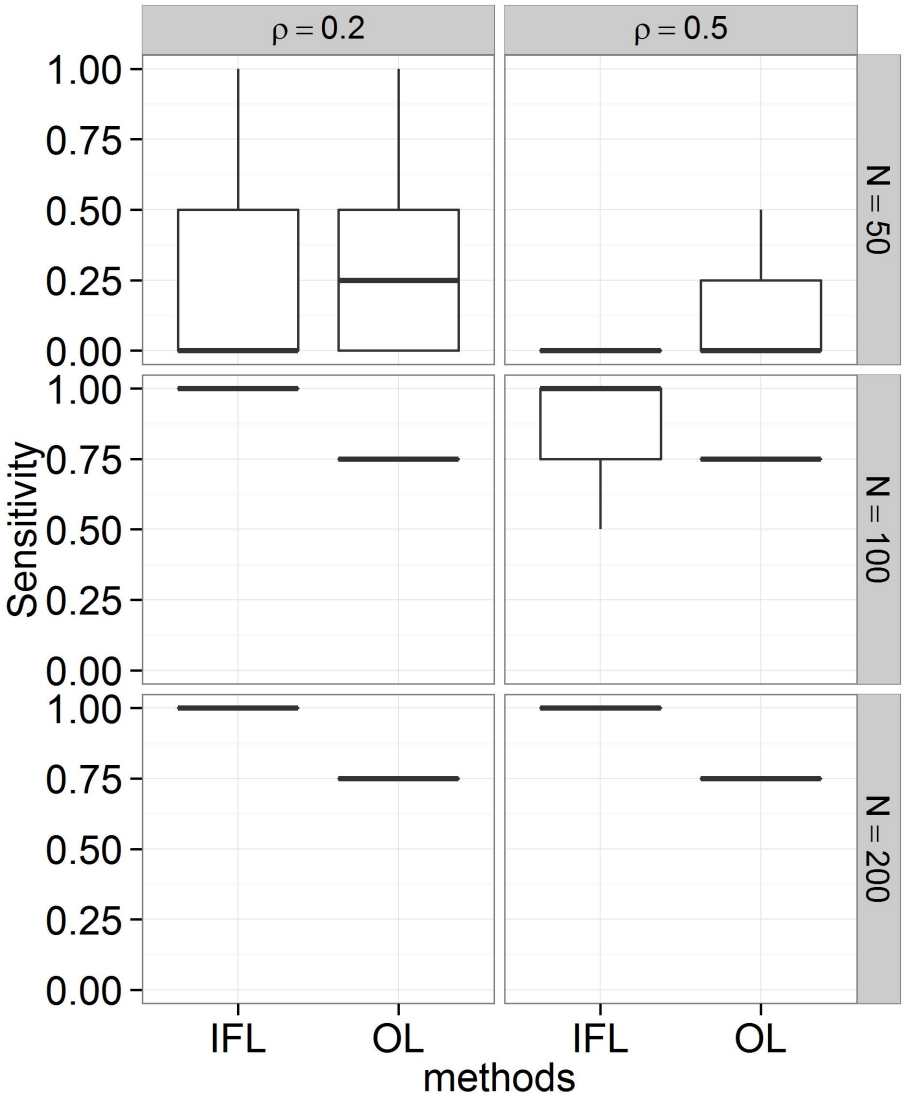

Figure 7. Sensitivity when the main effects are large (Scenario (iii) and (iv), Survival).

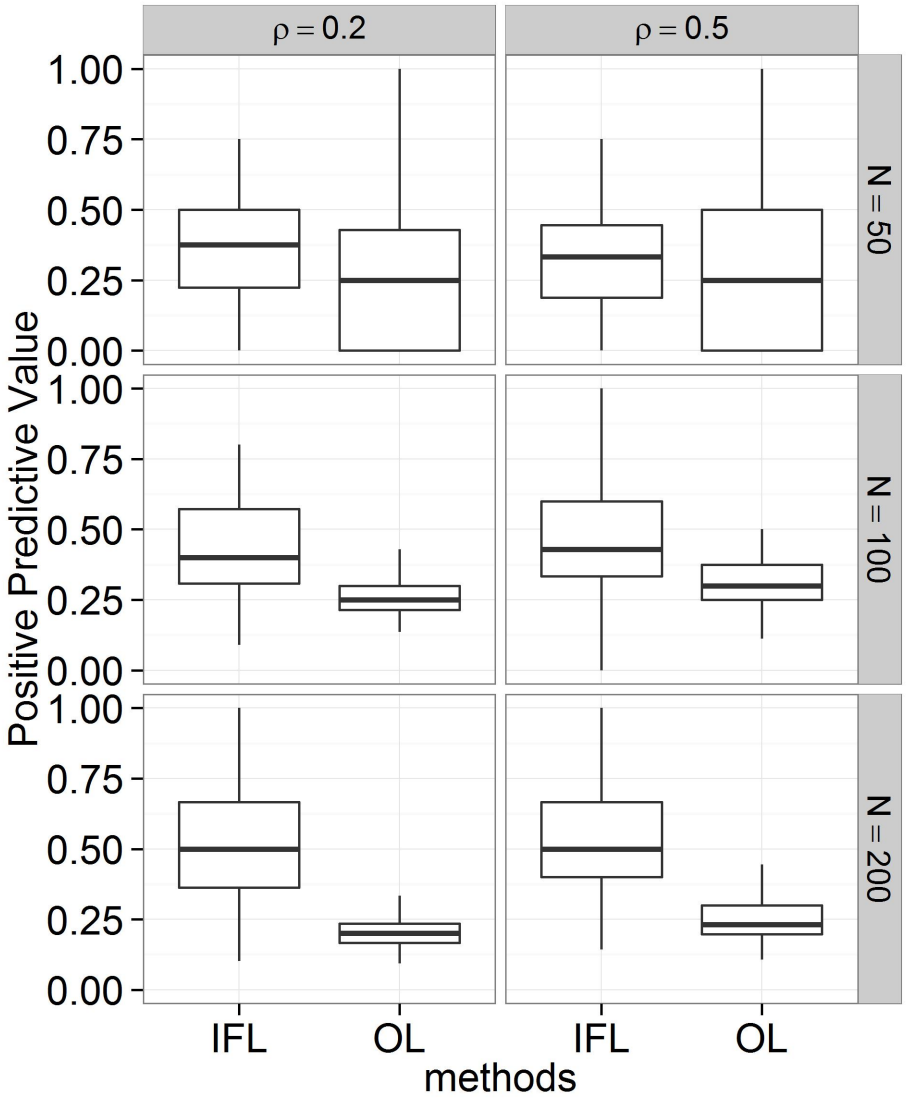

Figure 8. PPV when the main effects are large (Scenario (iii) and (iv), Survival).

2.6.3.2 Type I Error

The same model and data generating procedure are used except that the interaction coefficients $\gamma's$ are all set to 0. In total 500 runs of simulation are conducted for every case. The proportion of times that at least one interaction covariate is selected is calculated as the estimated Type I Error for each method.

Table II shows the result for the cases when the main effects are relatively large. The cases with moderate main effects have similar results. From Table II, we can see that the Type I Error of our method is much smaller than that of the OL method. When the sample size increases from 50 to 200, our method improves greatly while the OL method does not improve at all.

2.6.4 Summary

In this chapter, we introduced an interaction filtering procedure that can select relevant interactions without modeling the main effects. The proposed method works for both the continuous and the survival response, and is shown to be a reasonable procedure not relying on model specifications. Simulation is conducted to compare the performance of the proposed method and the ordinary multivariate regression with LASSO. Overall, our method shows a better performance in Sensitivity, Positive Predictive Value and Type I Error, especially when the sample size is moderately large.

TABLE II

TYPE I ERROR (LARGE MAIN EFFECTS, SURVIVAL)

p	N	Method	Type I Error
0.2	50	IFL	0.166
		OL	0.924
	100	IFL	0.142
		OL	0.996
	200	IFL	0.062
		OL	1.000
0.5	50	IFL	0.178
		OL	0.944
	100	IFL	0.092
		OL	0.998
	200	IFL	0.034
		OL	1.000

CHAPTER 3

SIGNATURE BUILDING WITH QUANTITATIVE CRITERIA

In this chapter, we propose a subgroup signature building method based on quantitative criteria. The proposed method does not rely on a specific model, and works for both the continuous and the survival response.

3.1 Motivation

Some subgroup identification methods focus on prediction, either predicting a response or a score, and select desired patients based on the prediction result. These methods usually produce a clear quantitative description of patients' potential outcome before selection, which brings several advantages. First, researchers can have some sense about the characteristics of potential subgroups. Second, the goal of the subgroup identification procedure can be based on some quantitative feature. Third, this also helps make the procedure look less like a fishing adventure though it still is. But the performance of these methods relies on the prediction accuracy at individual level, which is actually a more difficult problem than simply selecting a subgroup.

Some other methods tries to maximize the interaction effect or treatment effect in a subgroup, based on either p values of the interaction terms or the difference between test statistics of subgroups. These methods focus directly on splitting the dataset, and is thus more optimized for subgroup identification rather than prediction. The training result can often provide well

separated subgroups in terms of drug effects. However, there are also several disadvantages. First, maximizing the effect does not provide enough information about the potential subgroup. One does not know what the maximized effect is until a subgroup is obtained. Second, it is difficult to adjust the subgroup identification goal as needed. One could end up with a very small group with a very large effect size. But since the whole procedure is built on maximizing the effect, you cannot do much about it. Third, the training result is often overly optimistic.

It is desirable to have a method that focuses directly on subgroup identification rather than prediction, but meanwhile also provides quantitative information about the patients being selected. It is also desirable that the method does not rely on any specific model, and can have an adjustable identification goal.

3.2 Setup and Notations

The same setup and notations are used as in Chapter 2.Howerver, to make this chapter self-contained, we repeat them here. Suppose we have a randomized clinical trial with two arms, in which patients are allocated at random to receive either the treatment or the control with some prespecified probabilities. The control could be a placebo or some other existing treatment. Let $Y \in R^1$ be the response variable and $T \in R^1$ be the binary treatment variable, where $T = 1$ and -1 correspond to the treatment arm and the control arm respectively. Let $X \in R^p$ be the p-dimensional random vector of baseline covariates, including an intercept covariate 1. Since it is a randomized trial, T and X are independent. Assume the observed dataset consists of N patients, and every patient data in the dataset is an independent and identically distributed copy of (Y, T, X).

3.3 Method Derivation for the Continuous Response

A natural question people would ask when taking a medicine is "how much can I improve?". This question is a perfect quantitative criterion for subgroup identification. Translating it to a mathematical formula, we have

$$E\big(Y \mid X, T = 1\big) - E\big(Y \mid X, T = -1\big) \geq 2d \tag{3.1}$$

where d is a pre-specified constant, and serves as a threshold for the desired subgroup or say signature-positive group ($Sig.+$). Patients satisfying Condition (3.1) are those we want to identify and put in the signature-positive group. The rest of the patients are automatically put in the signature-negative group ($Sig.-$).

Based on (3.1), a patient in the positive group should satisfy

$$E\big(Y - Td \mid X, T = 1\big) - E\big(Y - Td \mid X, T = -1\big) \geq 0 \tag{3.2}$$

Let $D_0(X)$ be the true or say correct subgroup identifier, where

$$D_0(X) = \begin{cases} 1 & \text{when belonging to } Sig.+ \text{ group} \\ -1 & \text{when belonging to } Sig.- \text{ group} \end{cases}$$

Note that every patient in the Sig.+ group should have a nonnegative value in (3.2), while every patient in the Sig.− group should have a negative value. Thus, $D_0(X)$ has to satisfy

$$D_0(x) \in \underset{D}{\operatorname{argmax}} \int \Big[E\big(Y{-}Td \mid X, T=1\big){-}E\big(Y{-}Td \mid X, T=-1\big)\Big] I\big(D(x)=1\big)\omega(x)\, dP(x) \quad (3.3)$$

where $\omega(x) \geq 0$ is a weight function depending on X, $P(x)$ is the distribution of X and $I(\cdot)$ is an indicator function.

We rewrite (3.3) as follows

$$\begin{aligned}
& \underset{D}{\operatorname{argmax}} \left\{ \int E\big(Y-Td \mid X, T=1\big)\big(1-I\left(D(x)=-1\right)\big)\omega(x)\, dP(x) \right. \\
& \qquad \left. - \int E\big(Y-Td \mid X, T=-1\big) I\big(D(x)=1\big)\omega(x)\, dP(x) \right\} \\
\Longleftrightarrow \quad & \underset{D}{\operatorname{argmin}} \left\{ \int E\big(Y-Td \mid X, T=1\big) I\big(D(x)=-1\big)\omega(x)\, dP(x) \right. \\
& \qquad \left. + \int E\big(Y-Td \mid X, T=-1\big) I\big(D(x)=1\big)\omega(x)\, dP(x) \right\} \\
\Longleftrightarrow \quad & \underset{D}{\operatorname{argmin}} \int \Big\{ E\big[\big(Y-Td\big) I\big(D(X)\neq T\big) \mid X, T=1\big] \\
& \qquad + E\big[\big(Y-Td\big) I\big(D(X)\neq T\big) \mid X, T=-1\big] \Big\} \omega(x)\, dP(x) \\
\Longleftrightarrow \quad & \underset{D}{\operatorname{argmin}} \int \left\{ E\left[\frac{\big(Y-Td\big) I\big(D(X)\neq T\big)}{TP(T=1)+(1-T)/2} \mid X, T=1 \right] \cdot P(T=1) \right. \\
& \qquad \left. + E\left[\frac{\big(Y-Td\big) I\big(D(X)\neq T\big)}{TP(T=1)+(1-T)/2} \mid X, T=-1 \right] \cdot P(T=-1) \right\} \omega(x)\, dP(x) \\
\Longleftrightarrow \quad & \underset{D}{\operatorname{argmin}} \int E\left[\frac{\big(Y-Td\big) I\big(D(X)\neq T\big)}{TP(T=1)+(1-T)/2} \mid X \right] \omega(x)\, dP(x) \\
\Longleftrightarrow \quad & \underset{D}{\operatorname{argmin}}\, E\left[\frac{\big(Y-Td\big) I\big(D(X)\neq T\big)}{TP(T=1)+(1-T)/2} \cdot \omega(X) \right] \qquad (3.4)
\end{aligned}$$

Note that $D(X)$ can be equivalently written as $sign\big(f(X)\big)$ for some function f, where

$$sign(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ -1 & \text{if } u < 0 \end{cases}$$

Thus, the objective function in (3.4) becomes

$$\min_f E\left[\frac{(Y - Td) \cdot I\Big(sign\big(f(X)\big) \neq T\Big) \cdot \omega(X)}{TP(T=1) + (1-T)/2}\right] \tag{3.5}$$

The indicator function in (3.5) is not easy to handle during the optimization process. In practice, we often use a convex surrogate loss function instead. By rewriting (3.5) as

$$\min_f E\left[\frac{(Y - Td) \cdot \omega(X)}{TP(T=1) + (1-T)/2} \cdot I\Big(f(X) \cdot T \leq 0\Big)\right], \tag{3.6}$$

using log-loss $l(u) = \log(1 + e^{-u})$ to replace the 0-1 loss $l_0(u) = I(u \leq 0)$, the objective function becomes

$$\min_f E\left[\frac{(Y - Td) \cdot \omega(X)}{TP(T=1) + (1-T)/2} \cdot \log\big(1 + e^{-f(X)T}\big)\right]. \tag{3.7}$$

Let $f^*(X)$ be the minimizer of (3.7). The corresponding subgroup identifier is then $D^*(X) = sign\big(f^*(X)\big)$.

The subgroup signature building problem is now transformed to a minimization problem. Clearly, the objective function (3.7) focuses on identification rather than prediction. Meanwhile, it is also based on an adjustable quantitative criterion: the response difference of before and

after taking the medicine needs to be greater than a threshold. Moreover, it does not rely on any specific models.

Knowing the actual distributions of the variables in (3.7) and having a corresponding formula for the expectation are usually very difficult if not impossible. Instead, we may approximate the expectation with some observed data and seek to minimize the following empirical version of the objective function

$$\min_{f} \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i - T_i d}{T_i P(T=1) + (1-T_i)/2} \cdot \log\left(1 + e^{-f(X_i)T_i}\right) \cdot \omega(X_i) \tag{3.8}$$

The minimizer $\hat{f}^*(X)$ of (3.8) is thus an approximation of $f^*(X)$, and the corresponding $\hat{D}^*(X) = sign\left(\hat{f}^*(X)\right)$ is an approximation of $D^*(X)$

3.4 Fisher Consistency

After a series of derivation, we have the objective function (3.7) and its empirical version (3.8). We need to show that the method works as expected, i.e., $\hat{D}^*(X)$ is Fisher consistent. More specifically, we show the following result

Theorem 1. Define

$$R(f) = E\left[\frac{(Y - Td)\omega(X)}{TP(T=1) + (1-T)/2} \cdot \log\left(1 + e^{-f(X)T}\right)\right]$$

If function $f^*(X)$ minimizes $R(f)$, then $D^*(X) = D_0(X)$ almost surely, where $D_0(X)$ is the true subgroup identifier satisfying

$$D_0(X) = \begin{cases} 1 & \text{if } E(Y \mid X, T=1) - E(Y \mid X, T=-1) \geq 2d \\ -1 & \text{if } E(Y \mid X, T=1) - E(Y \mid X, T=-1) < 2d \end{cases}$$

Proof. If $f^*(X)$ minimizes $R(f)$, then $f^*(X)$ minimizes $R(f \mid X)$ almost surely with respect to X. It suffices to show that if $f^*(X)$ minimizes $R(f \mid X)$, then $D^*(X) = D_0(X)$.

$$\begin{aligned} R(f \mid X) &= E\left[\frac{(Y-d)\,\omega(X)}{P(T=1)} \cdot \log\left(1+e^{-f(X)}\right) \mid X, T=1\right] \cdot P(T=1) \\ &\quad + E\left[\frac{(Y+d)\,\omega(X)}{P(T=-1)} \cdot \log\left(1+e^{f(X)}\right) \mid X, T=-1\right] \cdot P(T=-1) \\ &= E\left[(Y-d)\cdot\omega(X)\cdot\log\left(1+e^{-f(X)}\right) \mid X, T=1\right] \\ &\quad + E\left[(Y+d)\cdot\omega(X)\cdot\log\left(1+e^{f(X)}\right) \mid X, T=-1\right] \\ &= E\big[Y-d \mid X, T=1\big]\cdot\omega(X)\cdot\log\left(1+e^{-f(X)}\right) \\ &\quad + E\big[Y+d \mid X, T=-1\big]\cdot\omega(X)\cdot\log\left(1+e^{f(X)}\right) \end{aligned} \tag{3.9}$$

To find the minimizer $f^*(X)$, simply take the derivative of (3.9) with respect to f and make it zero. Hence, we have

$$\begin{aligned} \frac{\partial R(f \mid X)}{\partial f} &= E\big[Y-d \mid X, T=1\big]\cdot\omega(X)\cdot\frac{-1}{1+e^f} + E\big[Y+d \mid X, T=-1\big]\cdot\omega(X)\cdot\frac{e^f}{1+e^f} \\ &= 0 \end{aligned}$$

$$\Longrightarrow f^*(X) = \log E\big[Y - d \mid X, T = 1\big] - \log E\big[Y + d \mid X, T = -1\big]$$

It is straightforward to see that this local minimum is also a global minimum. Clearly,

$$sign\big(f^*(X)\big) = \begin{cases} 1 & \text{if } E\big(Y \mid X, T = 1\big) - E\big(Y \mid X, T = -1\big) \geq 2d \\ -1 & \text{if } E\big(Y \mid X, T = 1\big) - E\big(Y \mid X, T = -1\big) < 2d \end{cases}$$

Note that $D^*(X)$ is defined as $sign\big(f^*(X)\big)$..

3.5 Extension to the Survival Response

The objective function (3.7) and its empirical version (3.8) do not rely on any specific model. It is thus straightforward to extend it to the survival response. The corresponding quantitative criterion for the survival response is that "the survival time difference between before and after taking the medicine needs to be greater than a threshold", which is as follows

$$E\left(\tilde{Y} \mid X, T = 1\right) - E\left(\tilde{Y} \mid X, T = -1\right) \geq 2d \tag{3.10}$$

where $\tilde{Y}$ is the survival time (Y is the observed time). However, unlike the continuous response case, we usually have censoring in the time to event data. We only observe Y. Some observed times are survival times, while others are censoring times. To address this issue, we need some further assumptions.

In Chapter 2, we assumed that the event indicator ξ is independent of T and X. In other words,

$$I\left(\tilde{Y}_c - \tilde{Y} \geq 0\right) \perp\!\!\!\perp X, T$$

where $\tilde{Y}_c$ is the censoring time. Hence, it is not unreasonable to move one step further and assume

$$\tilde{Y}_c - \tilde{Y} \perp\!\!\!\perp X, T$$

Based on the above independence assumption, we have

$$\begin{aligned} E(Y \mid X, T) &= E\left[\tilde{Y} \cdot \xi + \tilde{Y}_c \cdot (1-\xi) \mid X, T\right] \\ &= E\left[\tilde{Y} \cdot \xi \mid X, T\right] + E\left[\left(\tilde{Y}_c - \tilde{Y}\right) \cdot (1-\xi) \mid X, T\right] + E\left[\tilde{Y} \cdot (1-\xi) \mid X, T\right] \\ &= E\left[\tilde{Y} \mid X, T\right] + E\left[\left(\tilde{Y}_c - \tilde{Y}\right) \cdot (1-\xi)\right] \end{aligned}$$

Hence, the survival time difference between before and after taking the medicine becomes

$$\begin{aligned} & E\left[\tilde{Y} \mid X, T=1\right] - E\left[\tilde{Y} \mid X, T=-1\right] \\ = \; & E[Y \mid X, T=1] - E\left[\left(\tilde{Y}_c - \tilde{Y}\right) \cdot (1-\xi)\right] - E[Y \mid X, T=-1] + E\left[\left(\tilde{Y}_c - \tilde{Y}\right) \cdot (1-\xi)\right] \\ = \; & E[Y \mid X, T=1] - E[Y \mid X, T=-1] \end{aligned}$$

Thus, we can simply treat the observed time as the actual survival time, and perform the signature building procedure with the objective function (3.7) and (3.8) in the same way as in the continuous response case.

3.6 Large Sample Properties under Linear Function Setup

Objective functions (3.7) and (3.8) do not specify the form of function $f(X)$. In practice, we are interested in finding a linear function $f(X) = \beta' X$, $\beta \in R^p$, which is intuitive for interpretation and easy to use. The corresponding objective function and its empirical version are as follows

$$R_0(\beta) = E\left[\frac{(Y - Td)\,\omega(X)}{TP(T=1) + (1-T)/2} \cdot \log\left(1 + e^{-\beta' XT}\right)\right] \stackrel{\triangle}{=} E\left[g(\beta; X, Y, T)\right]$$

$$R_n(\beta) = \frac{1}{N}\sum_{i=1}^{N} \frac{(Y_i - T_i d)\,\omega(X_i)}{T_i P(T_i=1) + (1-T_i)/2} \cdot \log\left(1 + e^{-\beta' X_i T_i}\right) \stackrel{\triangle}{=} \frac{1}{N}\sum_{i=1}^{N} g(\beta; X_i, Y_i, T_i)$$

where

$$g(\beta; X, Y, T) = \frac{(Y - Td)\,\omega(X)}{TP(T=1) + (1-T)/2} \cdot \log\left(1 + e^{-\beta' XT}\right)$$

Note that shifting the response Y by some constant does not affect the quantitative criterion (3.1) and thus does not affect the true group identifier $D_0(X)$. As a result, we can always shift the response Y by some constant such that $(Y - Td)$ is positive. This condition is important for the implementation as well as the large sample properties.

Let $\beta_0 = \underset{\beta}{\operatorname{argmin}}\, R_0(\beta)$. Under the following regularity conditions:

(i) $R_0(\beta) < \infty$ for all $\beta \in R^p$

(ii) $E \parallel \frac{\partial g(\beta;X,Y,T)}{\partial \beta} \parallel^2 < \infty$ for each β in a neighborhood of β_0

(iii) $R_0(\beta)$ is twice differentiable at β_0

We will have

Theorem 2. $\beta_n \to \beta_0$ with probability 1, if β_0 is unique and the regularity condition(i) holds.

Proof. Using result in 3.2.2 of Boyd and Vandenberghe (34), $g(\beta; X, Y, T)$ is convex with respect to β for every fixed (X, Y, T). Also note that $\frac{(Y - Td)\omega(X)}{TP(T=1) + (1-T)/2} > 0$. Since β_0 is unique and the regularity condition(i) holds, it follows from Theorem 1 of Niemiro (35) that $\beta_n \to \beta_0$ with probability 1.

Theorem 3. $\sqrt{n}(\beta_n - \beta_0) \xrightarrow{D} N\left(0, H^{-1}VH^{-1}\right)$, if β_0 is unique and the regularity conditons (i), (ii), (iii) hold, where $H = \frac{\partial^2 R_0(\beta)}{\partial \beta^2}|\beta_0$, $V = Var\left[\frac{\partial g(\beta; X, Y, T)}{\partial \beta}|\beta_0\right]$.

Proof. $g(\beta; X, Y, T)$ is convex with respect to β for every fixed (X, Y, T) and $\frac{(Y - Td)\omega(X)}{TP(T=1) + (1-T)/2} > 0$. Since β_0 is unique and the regularity conditons (i), (ii), (iii) hold, the result follows directly from Theorem 4 of Niemiro (35).

One simple example satisfying these regularity conditions is the case when the support of X and Y are finite. Then the expectations are simply the weighted sum, and we can easily verify those conditions. In real world applications, all measurements are taken by some instrument

with limited range and precision, so it is not unreasonable to assume that these measurements have finite support.

3.7 Remarks for Implementation

In practice, (3.8) is the objective function that we actually deal with. As has been mentioned before, shifting the response Y by some constant does not affect the quantitative criterion (3.1) and thus does not affect the true group identifier $D_0(X)$. To make the objective function convex, we sometimes need to shift the response so that every $(Y_i - T_i d)$ is positive.

One direct way to optimize the twice differentiable convex function is to use the Newton-Raphson algorithm. However, the objective function (3.8) is based on log-loss which is essentially the loss function for the logistic regression. To see this, rewrite the likelihood of a logistic regression as follows

$$\begin{aligned} l &= \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right] \\ &= \sum_{y_i=1} \log \frac{1}{1 + e^{\beta' x_i (1-2y_i)}} + \sum_{y_i=0} \log \frac{1}{1 + e^{\beta' x_i (1-2y_i)}} \\ &= -\sum_{i=1}^{N} \log\left(1 + e^{\beta' x_i (1-2y_i)}\right) \end{aligned}$$

where $y_i = 0$ or 1 is the binary response, x_i is the q–dimensional covariates. p_i is the probability that Y_i equals 1. Clearly, the likelihood above has a similar form as (3.8). So, instead of writing a separate algorithm, we can directly use the weighted logistic regression to solve the problem, which makes it very convenient.

During the derivation, we assume $E(Y \mid X, T = 1) - E(Y \mid X, T = -1) \geq 2d$ is the quantitative criteriion for the signature positive group. Note that $T = 1$ is not necessarily the treatment arm and $T = -1$ is not necessarily the control arm. By switching the code of treatment and control, we can effectively change between preferring a larger response and a smaller response.

The threshold parameter d can also serve as a tuning parameter if we have no specific target subgroup in mind. In this case, a series of d values go through a cross-validation procedure and the p value of the interaction between the selected subgroup and the treatment is calculated for each d. The best d that leads to the smallest interaction p value is then selected. This procedure is quite like those methods that maximize the interaction effects, but the tuning parameter d provides a way to adjust the greediness and base the training procedure on the potential testing result.

One potential improvement to the current procedure is to include an L_1 penalty in the objective function such that the signature building step can also do variable selection. Then, the interaction filtering procedure introduced in Chapter 2 can work as a pre-filtering step with less stringent regulation.

3.8 Relationship with Outcome Weighted Learning Method

The outcome weighted learning method (OWL) tries to maximize the expected reward given that the treatment assignment rule is implemented, i.e., $\max E[Y \mid T = D(X)]$. Their derived objective function is as follows

$$\min_{D} E\left[\frac{Y \cdot I(T \neq D(X))}{TP(T = 1) + (1 - T)/2}\right] \tag{3.11}$$

Compared with (3.4), the above objective function is a special case when $d = 0$ and $\omega(X) = 1$. However, the starting point of the two methods are completely different. It is therefore interesting to see whether maximizing $E\left[Y \mid T = D(X)\right]$ is actually equivalent to our quantitative criterion with $d = 0$.

According to our quantitative criterion, when d equals 0, patients should take the treatment (i.e., belong to the Sig.+ group) as long as it is no worse than the control for them, and should take the control otherwise. In other words, patients should take the medicine that they can benefit from. This is a practical case when two competing drugs are available. Though maximizing $E\left[Y \mid T = D(X)\right]$ seems doing the same thing, they are actually different. To see this, we provide a simple data example as follows:

TABLE III

ILLUSTRATIVE DATA EXAMPLE

Patient ID	Covariate X	Treatment T	Response Y
1	1	-1	4
2	1	1	0
3	1	1	0
4	2	-1	0
5	2	1	1
6	2	1	1
7	3	-1	0
8	3	1	4
9	3	1	4

In this illustrative data example, we are lucky to have patients with the same covariate value in both arms, and we assume the response has no variance. The benefits of taking the treatment are -4, 1 and 4 for $X = 1$, 2 and 3 respectively. Let us consider two options for subgroup selection. Option 1, $\mathtt{Sig.-group} = \{X = 1 \text{ or } 2\}$ and $\mathtt{Sig.+group} = \{X = 3\}$; Option 2, $\mathtt{Sig.-group} = \{X = 1\}$ and $\mathtt{Sig+.group} = \{X = 2 \text{ or } 3\}$. Clearly, when $d = 0$, our method will choose Option 2 which gives patients the medicine that they can benefit from. However, if we calculate $E\left[Y \mid T = D(X)\right]$, we have 3 and 2.8 for Option 1 and Option 2 respectively. Maximizing $E\left[Y \mid T = D(X)\right]$ will thus lead to Option 1, which is not a reasonable choice.

Actually, maximizing $E\left[Y \mid T = D(X)\right]$ is equivalent to our quantitative criterion with d=0 only when $P(T = 1) = P(T = -1) = 1/2$. It is not a good objective function in general. Though the OWL method ends up with (3.11), which is a special case of our method, its starting point is not correct.

3.9 Simulation Studies

3.9.1 Simulation Setup

In this section, we perform a series of studies to investigate the performance of the proposed subgroup signature building method with quantitative criteria (QC). We want to see if the QC method can successfully select the patients who belong to the Sig.+ group when the true interaction covariates are provided. More specifically, the following metrics are considered

(i) Sensitivity

Definition The probability that a patient is selected by the QC method given that the patient is truly in the Sig.+ group.

Calculation

$$\frac{\left|\left\{\mathtt{Patient}_i : \mathtt{Patient}_i \text{ is in the } \mathtt{Sig.+} \text{ group } \& \ \mathtt{Patient}_i \text{ is selected}\right\}\right|}{\left|\left\{\mathtt{Patient}_i : \mathtt{Patient}_i \text{ is in the } \mathtt{Sig.+} \text{ group }\right\}\right|}$$

(ii) Positive Predictive Value (PPV)

Definition The probability that a selected patient is truly in the Sig.+ group.

Calculation

$$\frac{\left|\left\{\mathtt{Patient}_i : \mathtt{Patient}_i \text{ is in the } \mathtt{Sig.+} \text{ group } \& \ \mathtt{Patient}_i \text{ is selected}\right\}\right|}{\left|\left\{\mathtt{Patient}_i : \mathtt{Patient}_i \text{ is selected }\right\}\right|}$$

We use a similar data generating procedure as in Chapter 2. To make this section self-contained, we repeat it here. For the continuous response, the following model is used to generate the dataset

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon$$

where p is set to be 100, and the covariates $(X_1, X_2, ..., X_{100})$ follow a multivariate normal distribution with a compound symmetric variance-covariance matrix

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{100} \end{pmatrix} \sim N\left(\begin{pmatrix} 5 \\ 5 \\ \vdots \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}\right)$$

The means of the covariates are set to 5 to mimic the real world data which usually have positive measurements. The random error ϵ follows Normal (0, 1). The probability that a patient belongs to the treatment arm is 2/3, i.e., $P(T = 1) = 2/3$, $P(T = -1) = 1/3$.The values of $\gamma's$ are as follows

$$\gamma_1 = 1, \gamma_2 = -1, \gamma_3 = 2, \gamma_4 = -2, \text{ and all other } \gamma = 0$$

The four true interaction covariates are therefore X_1, X_2, X_3 and X_4. The values of $\beta's$ are as follows

$$\beta_5 = 2, \beta_6 = -2, \beta_7 = 4, \beta_8 = -4, \text{ and all other } \beta = 0$$

The four main effects are relatively large compared with the interaction effects.

For the survival response, we generate the data from the following model

$$Y = exp\left\{0.1 \cdot \left(\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon\right)\right\}$$

All the settings and parameters in the survival response case are exactly the same as in the continuous response case. The event indicator is randomly generated to induce 20% censoring rate.

We set the desired subgroup to have a threshold parameter $d = 1$ and 0.1 for the continuous response case and the survival response case respectively, which accounts for about 35% of the study population. The group that a patient actually belongs to can be calculated through

comparing $E(Y \mid X, T = 1) - E(Y \mid X, T = -1)$ with $2d$. For the continuous response case, it is straightforward to obtain the true group for every subject using the data generating model. For the survival response case, $E(Y \mid X, T)$ can be calculated through the moment generating function of the normal distribution induced by $\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon$

For both the continuous and the survival response, the following scenarios are considered.

(i) $\rho = 0.2$, sample size $N = 50, 100, 200, 400$

(ii) $\rho = 0.5$, sample size $N = 50, 100, 200, 400$

Though we have 100 covariates in the generated dataset, here in this section, we are only interested in the performance of the QC method which is not responsible for variable selection. Therefore, we let the algorithm know that only X_1, X_2, X_3 and X_4 are the true interaction covariates.

We conducted 500 runs of simulation for every case mentioned above. In each run, a training dataset is randomly generated and the QC method is applied. Meanwhile, a testing dataset with a sample size of 5000 is also generated using the same parameter setting as the training dataset. The corresponding Sensitivity and Positive Predictive Value are calculated in the testing dataset for each run, and boxplots are generated to show the average performance.

3.9.2 Simulation Result for the Continuous Response

Figure 9 and Figure 10 are the boxplots of Sensitivity and Positive Predictive Value respectively for the continuous response case. When the sample size increases, both Sensitivity and Positive Predictive Value increase, and the variance of the two metrics from 500 runs decreases. The correlation among covariates has a slight negative effect on the performance.

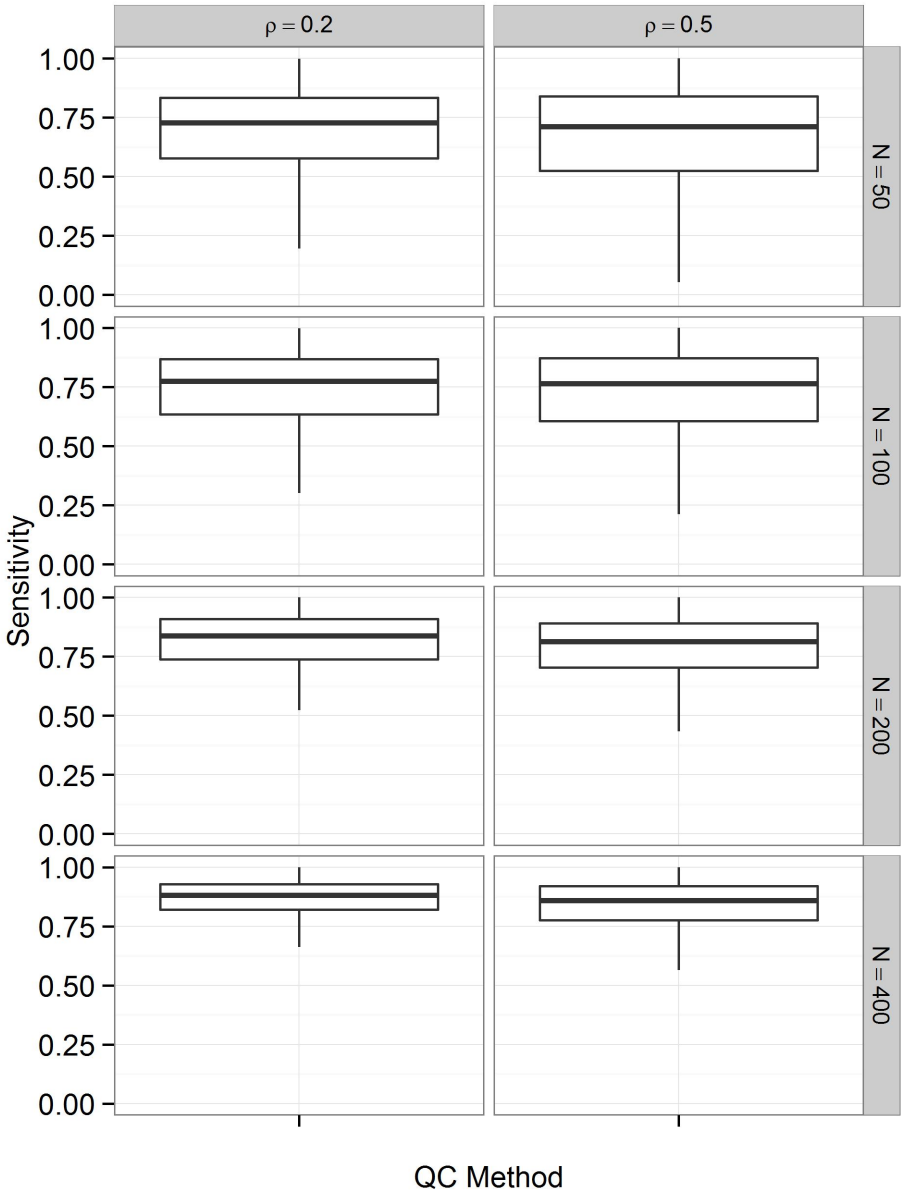

Figure 9. Sensitivity for the continuous response case (true interaction covariates known).

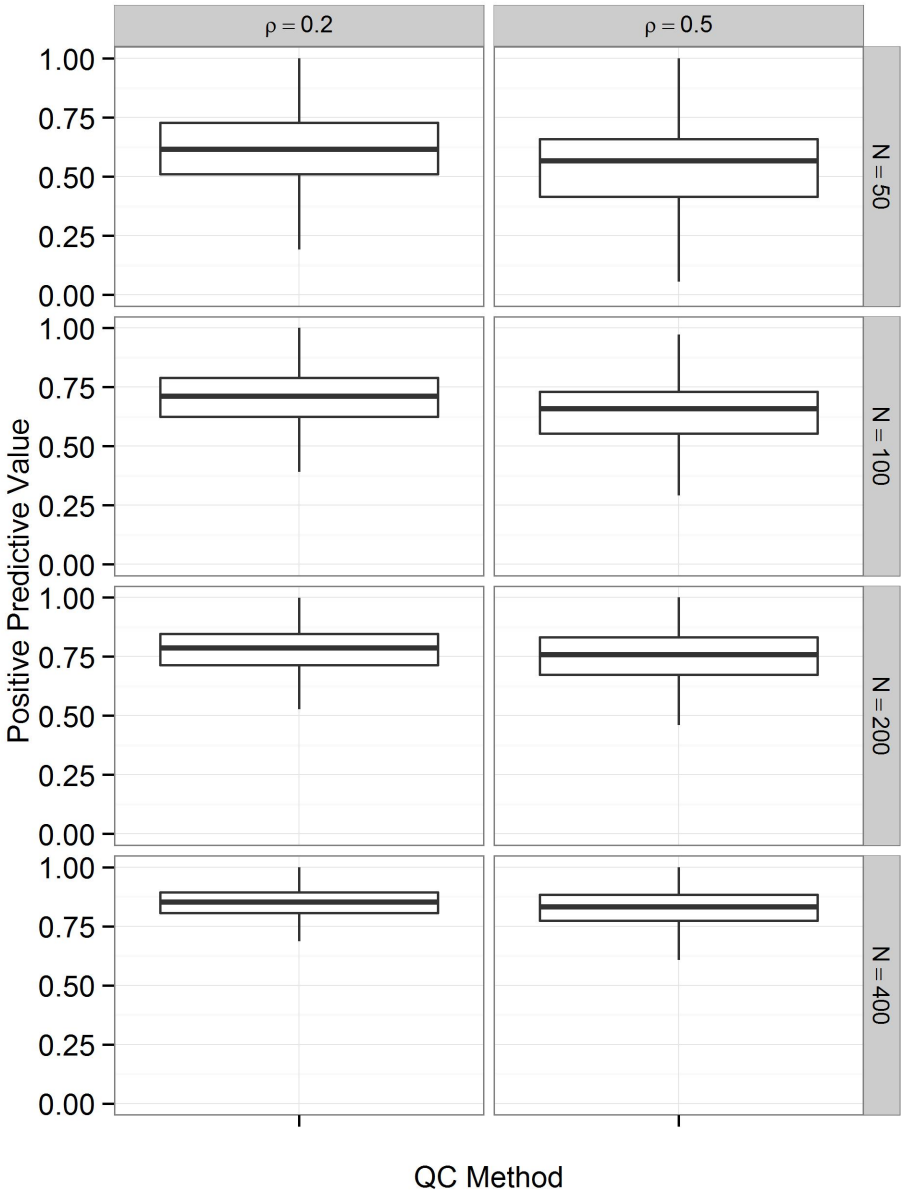

Figure 10. PPV for the continuous response case (true interaction covariates known).

3.9.3 Simulation Result for the Survival Response

Figure 11 and Figure 12 are the boxplots of Sensitivity and Positive Predictive Value respectively for the survival response case. The performance is not as good as the continuous response case, especially in terms of Positive Predictive Value. This is not surprising since the survival dataset is generated from a nonlinear model, which poses more challenge. But the pattern remains the same. When the sample size increases or the correlation among covariates decreases, both the Sensitivity and the Positive Predictive Value are improved.

3.10 Summary

In this chapter, we proposed a subgroup signature construction method based on an adjustable quantitative criterion. The proposed method does not rely on a specific model, and works for both the continuous and the survival response. The simulation result shows that the introduced method can successfully select the patients in the Sig.+ group with acceptable Sensitivity and Positive Predictive Value when the true interaction covariates are provided.

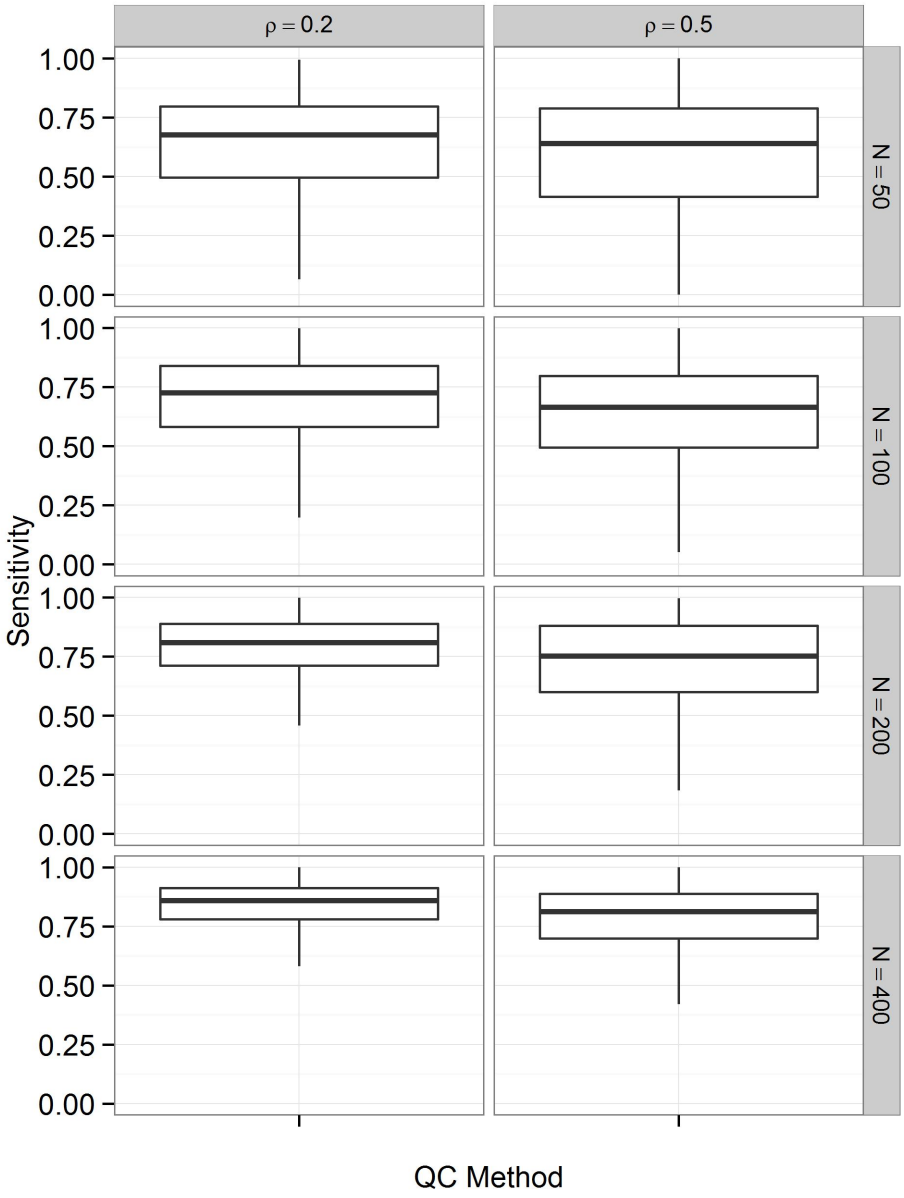

Figure 11. Sensitivity for the survival response case (true interaction covariates known).

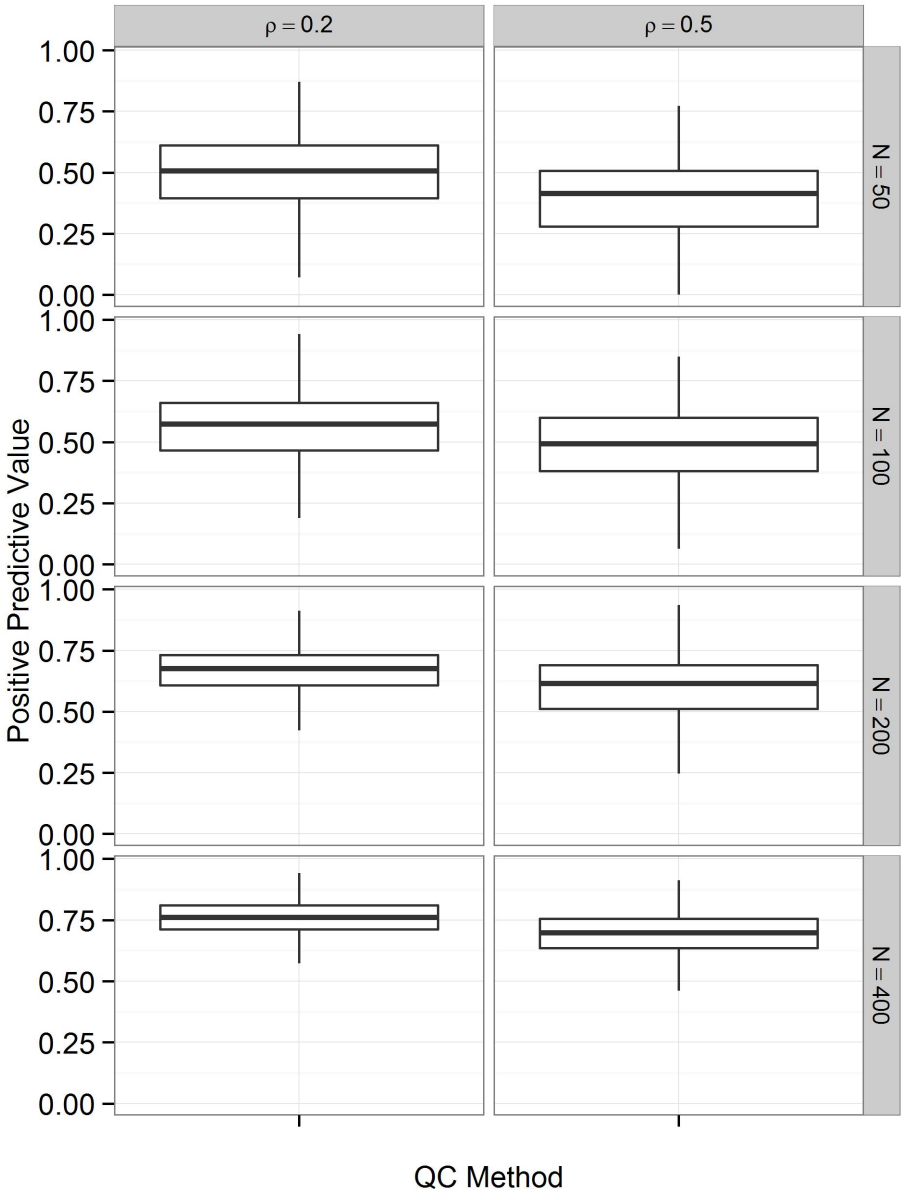

Figure 12. PPV for the survival response case (true interaction covariates known).

CHAPTER 4

PERFORMANCE OF THE COMBINED PROCEDURE

In this chapter, we combine the interaction filtering step (IFL) in Chapter 2 and the subgroup signature building step (QC) in Chapter 3 together into a complete subgroup identification procedure (IQ). The performance of the IQ method is first evaluated in a standalone manner and then will be compared with two other methods.

4.1 Sensitivity and Positive Predictive Value

In Chapter 3, we investigated the performance of the QC procedure when the true interaction covariates are known. In reality, there is no way to have this information, and we need to do variable selection. We repeat the exact same simulation as in Chapter 3, but the true interaction covariates are not assumed to be known. The IQ method will first select important interaction covariates through the IFL procedure and then the QC procedure will build an identification rule.

To make this chapter self-contained, we briefly repeat the simulation setup. For the continuous response, the following model is used to generate the dataset

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon$$

where p is set to be 100, and the covariates $(X_1, X_2, ..., X_{100})$ follow a multivariate normal distribution with a compound symmetric variance-covariance matrix

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{100} \end{pmatrix} \sim N\left(\begin{pmatrix} 5 \\ 5 \\ \vdots \\ 5 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \right)$$

The random error ϵ follows Normal (0, 1). The probability that a patients belongs to the treatment arm is $2/3$, i.e., $P(T = 1) = 2/3$, $P(T = -1) = 1/3$.The values of $\gamma's$ are as follows

$$\gamma_1 = 1, \gamma_2 = -1, \gamma_3 = 2, \gamma_4 = -2, \text{ and all other } \gamma = 0$$

The values of $\beta's$ are as follows

$$\beta_5 = 2, \beta_6 = -2, \beta_7 = 4, \beta_8 = -4, \text{ and all other } \beta = 0$$

For the survival response, we shall generate the data from the following model

$$Y = exp\left\{ 0.1 \cdot \left(\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \gamma_i X_i T_i + \epsilon \right) \right\}$$

All the settings and parameters in the survival response case are exactly the same as in the continuous response case. The event indicator is randomly generated to induce 20% censoring rate.

We set the desired subgroup to have a threshold parameter $d = 1$ and 0.1 for the continuous response case and the survival response case respectively, which accounts for around 35% of the study population.

For both the continuous and the survival response, the following scenarios are considered.

(i) $\rho = 0.2$, sample size $N = 50, 100, 200, 400$

(ii) $\rho = 0.5$, sample size $N = 50, 100, 200, 400$

We conducted 500 runs of simulation for every case mentioned above. Sensitivity and Positive Predictive Value, which have the same definition as in Chapter 3, are calculated in a testing dataset of 5000 patients for each run. Boxplots are generated to show the average performance.

Figure 13 and Figure 14 are the results for the continuous response case. Figure 15 and Figure 16 are the results for the survival response case. Clearly, the performance is not as good as when the true interaction covariates are known, but is still acceptable when the sample size is above 100. Note that the overall performance of the IQ method depends on both of the two sub-procedures, IFL and QC. If either of them goes extremely wrong, the result will not look good.

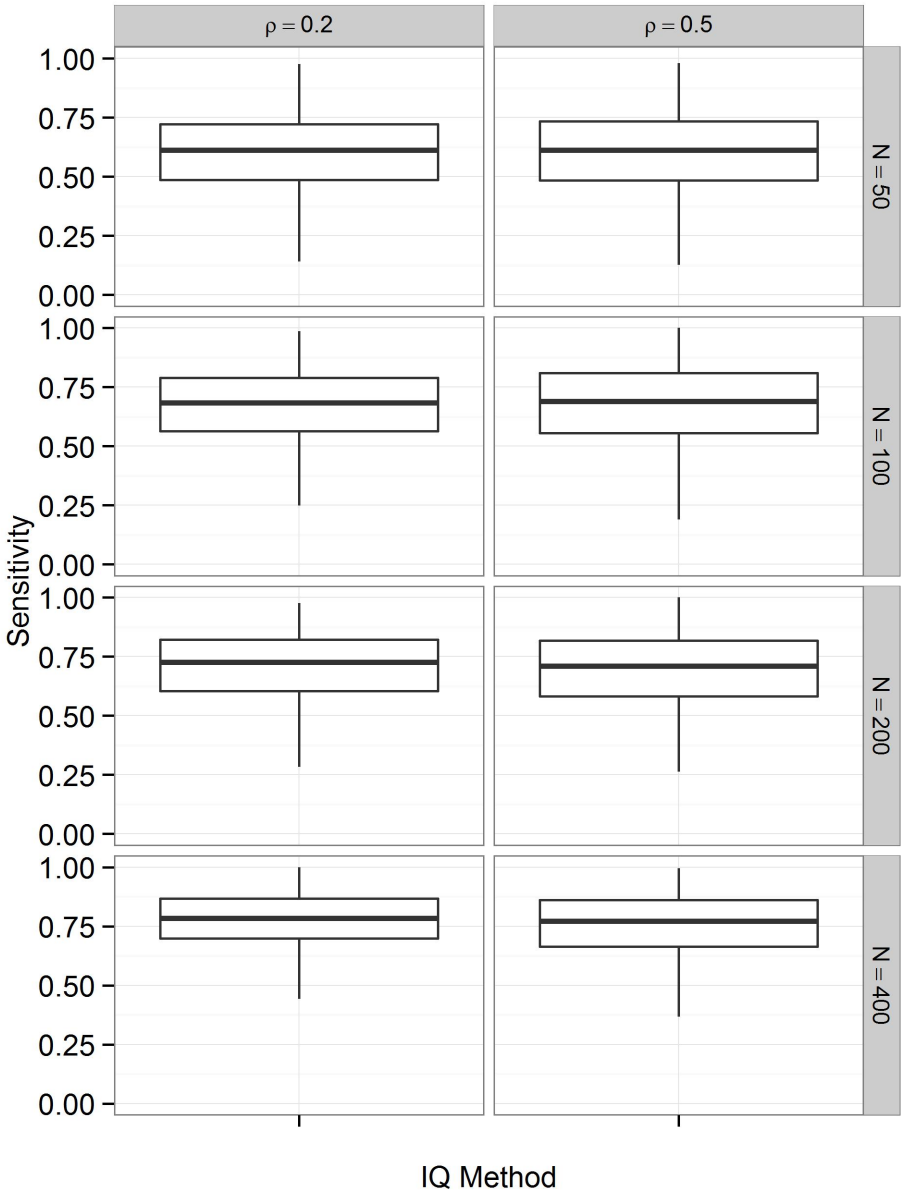

Figure 13. Sensitivity for the continuous response case (true interaction covariates not known).

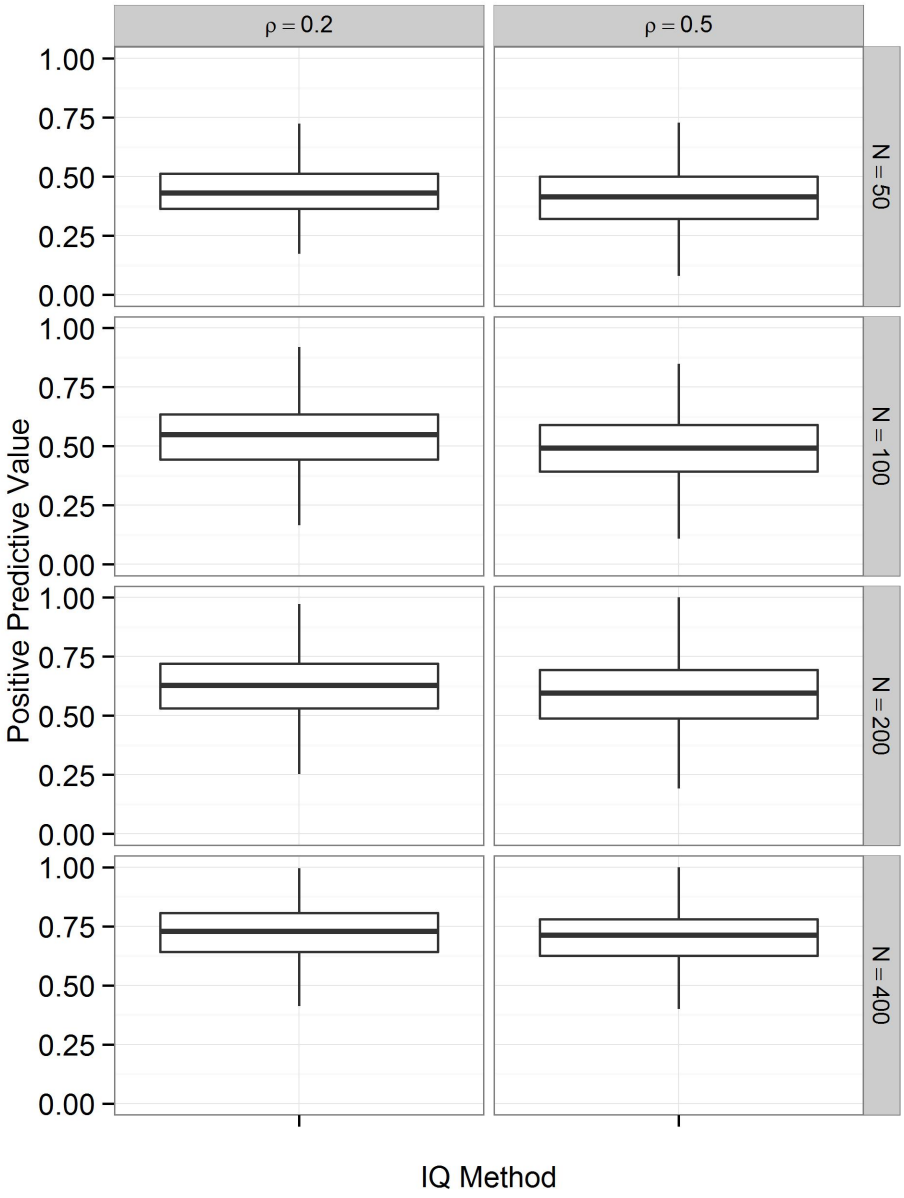

Figure 14. PPV for the continuous response case (true interaction covariates not known).

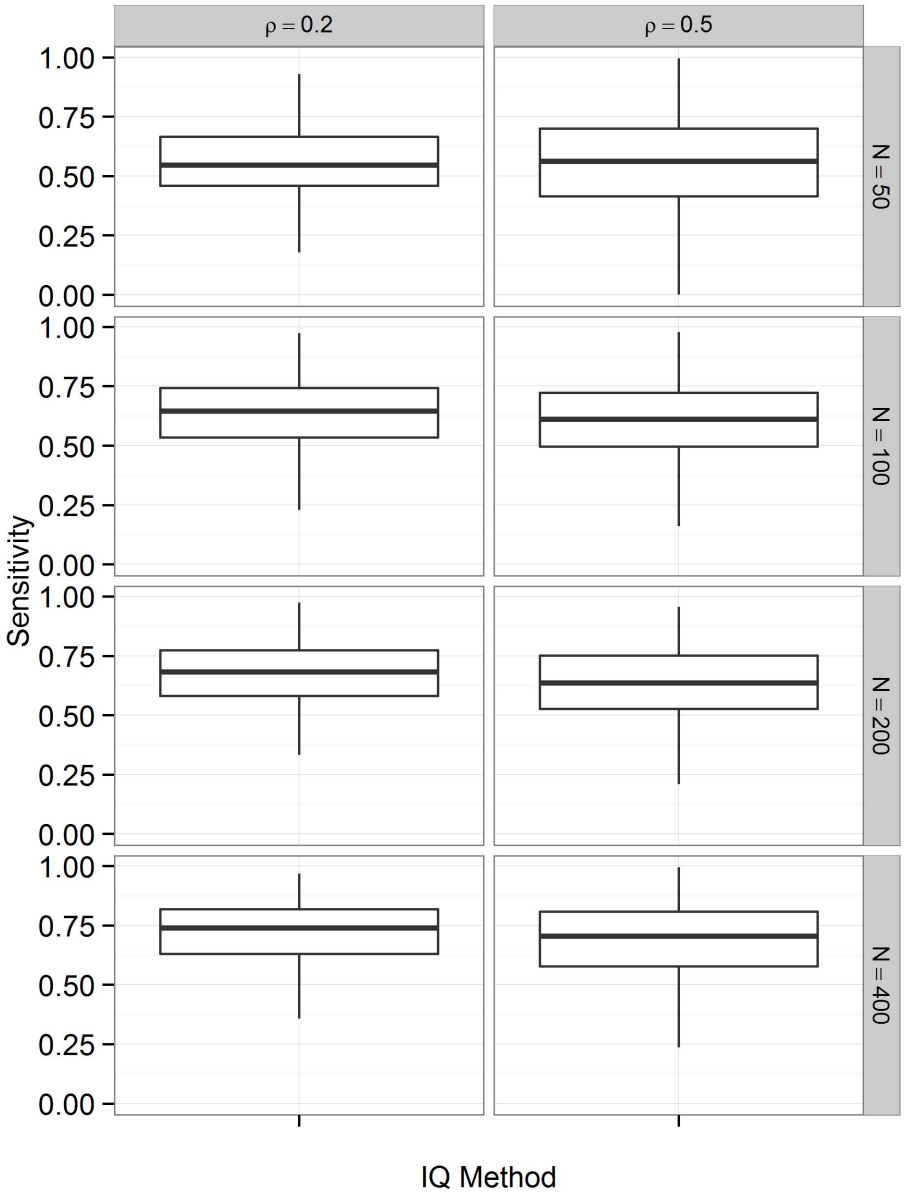

Figure 15. Sensitivity for the survival response case (true interaction covariates not known).

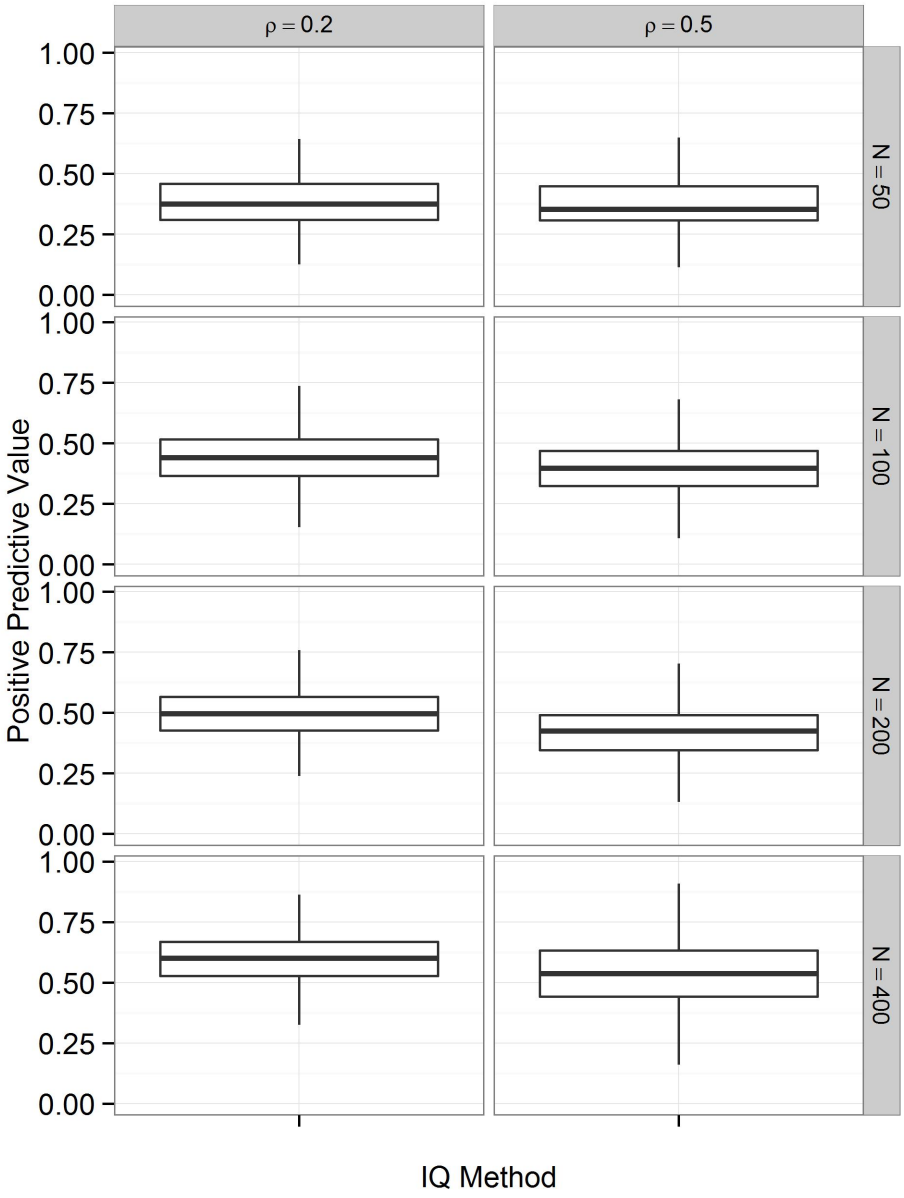

Figure 16. PPV for the survival response case (true interaction covariates not known).

4.2 Comparison with Two Other Methods

In this section, we will compare our IQ method with the AIM method (29) and the SIDES method (26). As we have mentioned in Chapter 1, AIM creates a score system which can be used to stratify patients. SIDES tries to find the subgroup with maximized treatment effect and build a collection of binary rules to select patients. Clearly, neither AIM nor SIDES can be used to select a subgroup with specified quantitative feature. To make the three methods comparable, we set the subgroup identification goal as finding a subgroup with the most significant improvement after taking the treatment, which is essentially what SIDES and many other methods do. But we have to modify the IQ method and the AIM method a little. For the IQ method, instead of setting a specific threshold parameter `d`, we treat `d` as a tuning parameter and let the algorithm decide the best `d` that leads to the most significant interaction effect (we mentioned this in Section 3.7). For the AIM method, after the score system is built, we look for a cutoff on the score such that patients above or below that cutoff will become the `Sig.+` group that has the most significant interaction effect.

We use the same simulation model as in Section 4.1. The correlation among the covariates is set to be 0.5. Since AIM and SIDES cannot handle high dimensional data, the number of covariates `p` is set to be 20. All other parameters remain the same as before. We generate a training set of 150 patients, and apply the learning results from the three methods on a testing set of 5000 patients.

Figure 17 depicts the training and testing result for the continuous case. Note that the vertical line segments are the error bars corresponding to one standard deviation, and the

values of N in the graph are the sample size of Sig.+ group or Sig.− group depending on its position. AIM is distracted by the main effects due to its incomplete model, and finds a subgroup with very strong prognostic effect but marginal predictive effect. More specifically, in both the training and the testing result of AIM, average response in the Sig.+ group is much larger than the average response in the Sig.− group, which indicates a strong prognostic effect or say main effect in the Sig.+ group. However, patients in the Sig.+ group only benefit a little more than the patients in the Sig.− group after taking the medicine. Both SIDES and IQ manage to select a subgroup that has very significant treatment effect. In the testing set, the average improvement in the two Sig.+ groups are about the same, but considering the sample size of the Sig.+ group and its variance (length of the error bars), IQ method is apparently better. This is mainly due to the fact that SIDES is greedy and tries to find the maximized effect in the training set, while the IQ method can adjust its greediness through the tuning threshold parameter d.

Figure 18 depicts the training and testing result for the survival case. Clearly, AIM is still distracted by the big main effects. As for SIDES and IQ, this is a clear example of why being greedy is a bad strategy. The training result of SIDES is much more significant than that of IQ, and the sample size distribution of the two are about the same. However, due to its non-adjustable greediness, the finding of SIDES is a false discovery. The pattern found in the training set disappears completely in the testing set. To the contrary, though the training result of IQ does not look so impressive, the pattern holds in the testing set.

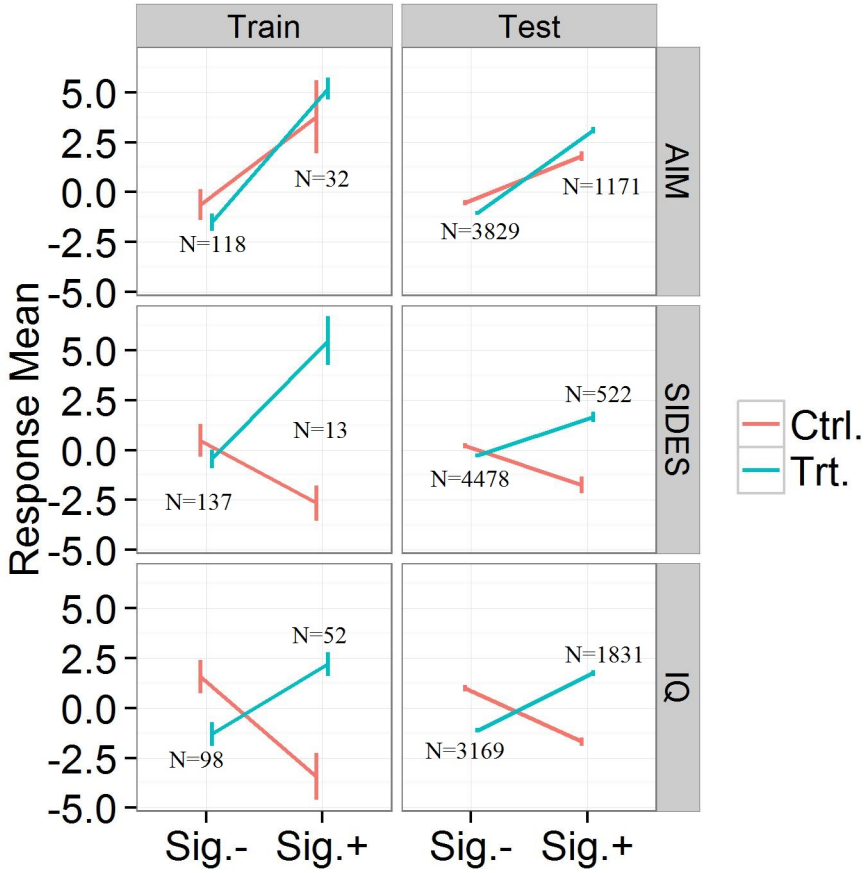

Figure 17. Comparison of methods for the continuous response case.

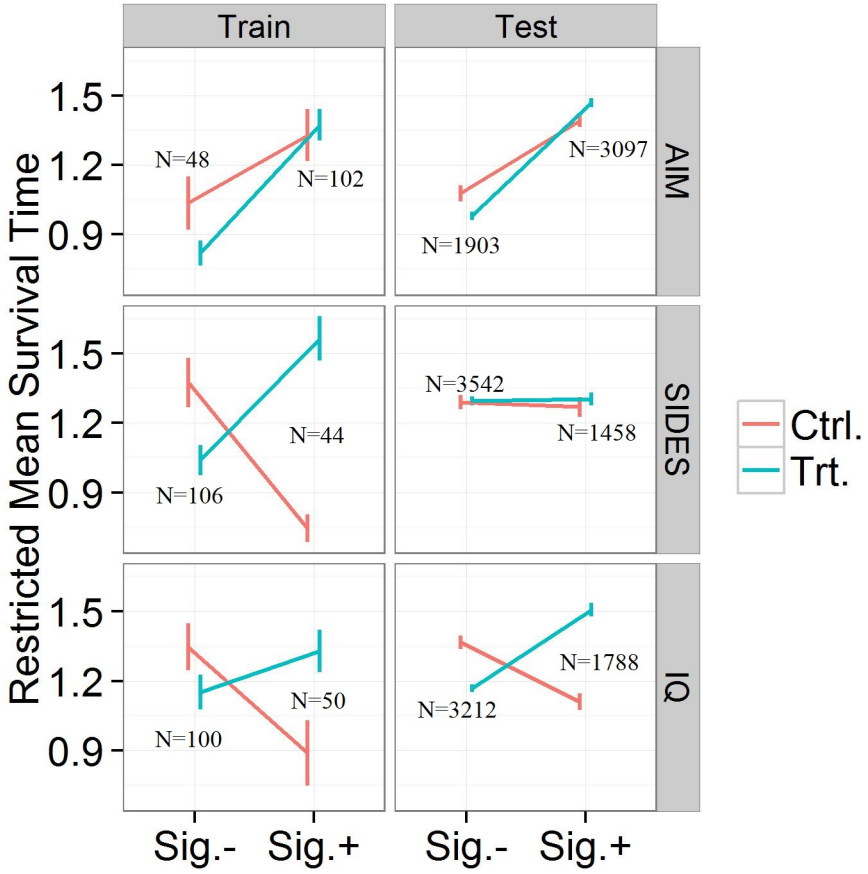

Figure 18. Comparison of methods for the survival response case.

CITED LITERATURE

1. Breiman, L., Friedman, J., Stone, C. J., and Olshen, R.: Classification and regression trees. New York, Chapman and Hall/CRC, 1984.

2. Butte, A.: The use and analysis of microarray data. Nature Reviews Drug Discovery, 1:951–960, 2002.

3. Furnival, G. and Wilson, R.: Regression by leaps and bounds. Technometrics, 16:499–511, 1974.

4. Hastie, T., Tibshirani, R., and Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction. Springer series in statistics, 2009.

5. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society : Series B (Methodological), 58:267–288, 1996.

6. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, page 11371143, 1995.

7. Lipkovich, I. and Dmitrienko, A.: Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. Journal of Biopharmaceutical Statistics, 24:130–153, 2014.

8. Chen, G., Zhong, H., Belousov, A., and Devanarayan, V.: A prim approach to predictivesignature development for patient stratification. Statistics in Medicine, 34:317–342, 2015.

9. Woodcock, J.: The prospects for "personalized medicine" in drug development and drug therapy. Clinical Pharmacology and Therapeutics, 81:164–169, 2007.

10. Mok, T. S., Ichinose, Y., Nishiwaki, Y., Ohe, Y., Chewaskulyong, B., Jiang, H., Duffield, E. L., Watkins, C. L., Armour, A. A., Fukuoka, M., Wu, Y., Thongprasert, S., Yang, J., Yang, C., Chu, D., Saijo, N., Sunpaweravong, P., Han, B., and Margono, B.: Gefitinib or carboplatinpaclitaxel in pulmonary adenocarcinoma. The New England Journal of Medicine, 361:947–957, 2009.

11. Slamon, D. J., Pegram, M., Baselga, J., Norton, L., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., and Wolter, J.: Use of chemotherapy plus a monoclonal antibody against her2 for metastatic breast cancer that overexpresses her2. The New England Journal of Medicine, 344:783–792, 2001.

12. Paez, J. G., Boggon, T. J., Naoki, K., Sasaki, H., Fujii, Y., Eck, M. J., Sellers, W. R., Johnson, B. E., Meyerson, M., Jnne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., and Lindeman, N.: Egfr mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science, 304:1497–1500, 2004.

13. Amado, R. G., Radinsky, R., Patterson, S. D., Chang, D. D., Wolf, M., Peeters, M., Cutsem, E. V., Siena, S., Freeman, D. J., Juan, T., Sikorski, R., and Suggs, S.: Wild-type kras is required for panitumumab efficacy in patients with metastatic colorectal cancer. Journal of Clinical Oncology, 26:1626–1634, 2008.

14. Song, Y. and Chi, G. H.: A method for testing a prespecified subgroup in clinical trials. Statistics in Medicine, 26:3535–3549, 2007.

15. Alosh, M. and Huque, M.: A flexible strategy for testing subgroups and overall population. Statistics in Medicine, 28:3–23, 2009.

16. Beckman, R., Clark, J., and Chen, C.: Integrating predictive biomarkers and classifiers into oncology clinical development programmes. Nature Reviews Drug Discovery, 10:735–748, 2011.

17. Food and Drug Administration, .: Enrichment strategies for clinical trials to support approval of human drugs and biological products. 2012.

18. Bonetti, M. and Gelber, R. D.: Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics, 5:465–481, 2004.

19. Sauerbrei, W., Royston, P., and Zapien, K.: Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches. Computational Statistics and Data Analysis, 51:4054–4063, 2007.

20. Bayman, E. O., Chaloner, K., and Cowles, M. K.: Detecting qualitative interaction: a bayesian approach. Statistics in Medicine, 29:455–463, 2010.

21. Sivaganesan, S., Laud, P. W., and Mller, P.: A bayesian subgroup analysis with a zeroenriched polya urn scheme. Statistics in Medicine, 30:312–323, 2011.

22. Loh, W.: Regression trees with unbiased variable selection and interaction detection. Statistica Sinica, 12:361–386, 2002.

23. Foster, J. C., Taylor, J. M., and Ruberg, S. J.: Subgroup identification from randomized clinical trial data. Statistics in Medicine, 30:2867–2880, 2011.

24. Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S.: Interaction trees with censored survival data. The International Journal of Biostatistics, 4:2–26, 2008.

25. Su, X., Tsai, C., Wang, H., Nickerson, D. M., and Li, B.: Subgroup analysis via recursive partitioning. Journal of Machine Learning Research, 10:141–158, 2009.

26. Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G.: Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. Statistics in Medicine, 30:2601–2621, 2011.

27. Friedman, J. H. and Fisher, N. I.: Bump hunting in high-dimensional data. Statistics and Computing, 9:123–143, 1999.

28. Zhao, L., Tian, L., Cai, T., Claggett, B., and Wei, L. J.: Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association, 108:527–539, 2013.

29. Tian, L. and Tibshirani, R.: Adaptive index models for marker-based risk stratification. Biostatistics, 12:68–86, 2011.

30. Tian, L., Alizadeh, A. A., Gentles, A., and Tibshirani, R.: A simple method for detecting interactions between a treatment and a large number of covariates. Available at http://arxiv.org/pdf/1212.2995.pdf, 2012.

31. Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R.: Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107:1106–1118, 2012.

32. Cox, D. R.: Regression models and life tables. Journal of the Royal Statistical Society. Series B (Methodological)., 34:187–220, 1972.

33. Zou, H. and Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67:301–320, 2005.

34. Boyd, S. and Vandenberghe, L.: Convex optimization. Cambridge, Cambridge University Press, 2004.

35. Niemiro, W.: Asymptotics for m-estimators defined by convex minimization. The Annals of Statistics., 20:1514–1533, 1992.

VITA

Education

- Ph.D in Statistics, University of Illinois at Chicago, 2015
- M.S in Electrical Engineering, Fudan University, 2010
- B.S in Electrical Engineering, Fudan Univeristy, 2007

Working Experience

- Extern and Summer Intern, Abbvie, Chicago, IL, 2013-2015
- Summer Intern, Mayo Clinic, Rochester, MN, 2013

Publication

- Y. Sun, X. Chen and Q. Lu, "Section-based object tracking" *IEEE Asia Pacific Conf. on Postgraduate Research in Microelectronics and Electronics*, 2009
- Y. Sun, X. Chen and Q. Lu, "Robust kernel design for ssd tracker", *IEEE Youth Conf. on Information, Computing and Telecommunications*, 2009

Joint Inventor of the Following Patents(China)

- Publication Time: 12/30/2009. Publication No.: CN201374123. "Traffic Flow Detection System Based on Video Sequence"
- Publication Time:5/12/2010. Publication No.: CN201457305U. "Electronic Locks based on Image"