VARIABLE SELECTION FOR DECISION MAKING

by

Lacey L. Gunter

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2009

Doctoral Committee:

Professor Susan A. Murphy, Co-Chair
Associate Professor Ji Zhu, Co-Chair
Professor Satinder Singh Baveja
Associate Professor Kerby A. Shedden

First and foremost I dedicate this to my wonderful husband, Jacob, who stood by my side the many long years this took to complete. I genuinely appreciate your sacrifices and I am a better person because of your support. I also dedicate this to my wise and caring parents, Kenneth and Joyce Jones. You instilled in me a great love for learning and you were a wonderful example of the values and principles you taught. You always made me believe I could accomplish anything I set my mind to. Last, but not least, I dedicate this to my two beautiful sons, Samuel and Jeremiah, who have put up with endless hours at the computer, late nights, early mornings, multiple babysitters and one too many simulations. Your smiles, hugs and unconditional love gave me the courage to face another day.

ACKNOWLEDGEMENTS

I wish to thank everyone who made this thesis possible.

I deeply appreciate my advisors Susan Murphy and Ji Zhu for everything they have done. You helped me grow and develop from a student to a researcher. You spent so many hours just listening to ideas, just as many hours offering ideas and you have helped me gain insights I would have never gained on my own. Your push for excellence has helped me to recognize my capabilities exceed my expectations. Your passion for research and the advancement of statistics is truly inspiring.

I warmly thank my student colleges who helped me get through the day to day drudgery required to complete a thesis. You entertained me with many hours of stimulating conversation, you encouraged me on the days nothing seemed to be working, you gave me new ideas when my research seemed to be stuck and you shared a common goal and a piece of my life I will never forget.

I wish to thank the many teachers throughout my life who helped me acquire the knowledge needed to get to this point, especially all the mathematics and statistics teachers and professors who recognized what my talents were and helped to develop them.

Thank you to Martin Keller and the many investigators who conducted the trial 'A Comparison of Nefazodone, the Cognitive Behavioral-analysis System of Psychotherapy, and Their Combination for Treatment of Chronic Depression', for the use of their data and John Rush for several hours of guidance and technical support.

Thank you to all the staff and secretaries who seemed to have a wealth of knowledge about all the important degree and funding issues that no one else seemed to know about. I am grateful for your kind words and quiet support.

I thank God for all the many blessings I have received. And of course, thank you to all my family and friends for your love, encouragement, listening ear, and constant motivation.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Many researchers collect a wealth of information on subjects without knowing a-priori which information will be most relevant. Broad data collection can help compensate for unanticipated problems in scientific trials due to imbalances in randomization, dropout or noncompliance. Extensive data collection also allows for researchers to explore new ideas and discover unknown relationships. As the number of variables increases, however, sample size needs increase and it becomes more difficult to decipher true relationships from noise - a problem often referred to as the curse of dimensionality

Constant advancements in computing power and data storage capabilities ensure the amount of information collected will likely continue to grow with time. Due to these advances in data collection and storage, variable selection has become a popular topic of research in the field of machine learning.

Eliminating variables that either fail to contribute relevant information or contribute redundant information to the task of learning can be very beneficial. Variable selection can save money and time used to collect unessential information, reduce computation time and improve efficiency and stability. Also, models that use fewer variables are often easier to understand and interpret. Many techniques have been

developed for doing variable selection, ranging from simple to sophisticated. Most of these techniques, however, were designed for applications focusing on prediction or classification.

Applications that focus on decision making must also deal with variable selection. Decision making applications occur in many fields and are becoming more prevalent as the demand for evidence based decision making grows. In these applications the final goal is to choose actions that result in the best future outcome. Prediction of the response represents a first step in finding optimal decisions, but is not the underlying goal.

While variable selection techniques developed for prediction can and are used in applications focused on decisions making, they have important drawbacks. They often leave behind small but important interaction variables that are critical when the ultimate goal is optimal decision making rather than optimal prediction. These variables are important because they play a role in determining which action is best for different subsets of the population. The variables qualitatively interact with the action.

In this thesis we propose new methods for variable selection that are geared toward decision making applications. These methods seek to find the variables which play a role in determining the best action and improve the overall outcome resulting from the chosen actions. The new methods are tested against recommended tests for qualitative interactions and popular variable selection techniques for prediction.

The thesis is outlined as follows. In Chapters II and III we give a background and framework for decision making applications and detail the current state of variable selection for machine learning. In Chapters IV and V we present the new methods for variable selection in a single time point decision making problem and demonstrate

their ability on real and simulated data. We conclude the thesis with a discussion on how to expand these ideas to do variable selection for sequential decision making applications.

CHAPTER II

Decision Making Applications

Decision making applications occur in many fields of research. They are applied in various areas such as medical decision making and artificial intelligence. Much of this research has been done outside the field of statistics. This chapter introduces the common components and many important aspects of decision making applications. The chapter briefly discusses methods for estimating optimal decisions and concludes with a short discussion on the need for variable selection in these applications.

The goal in decision making applications is to make decisions that result in the most desirable final outcome. The components of a decision making application are observations, actions and responses. At each decision time point t, we obtain observations about a subject, X_t. This information is used to select an action, A_t. We then receive a response, R_t. The response variable is an unknown (possibly random) function of any prior actions, observations and patient outcomes subsequent to A_t. For example, an application with two decision time points would consist of trajectories of the form $(X_1, A_1, R_1, X_2, A_2, R_2)$. The observations and actions may be categorical or continuous, while the responses are assumed to be continuous. The observations, X_t, may be multi-dimensional and affected by previous actions. A policy or strategy, π, is a set of stochastic or deterministic decision rules mapping

the space of past observations and actions to the current action space. In other words, at time t, π_t defines the probability for choosing action, A_t, given the history at time t, $H_t = (X_1, A_1, X_2, A_2, ..., X_{t-1}, A_{t-1}, X_t)$. The responses, R_t, give us some indication of the desirability of the current action and/or strategy and are often referred to as rewards. Thus, the goal is to find the policy π^*, which optimizes the sum of the responses.

A simple example of a decision making application with one time point is a clinical trial to test two alternative treatments. In this case the observation vector consists of baseline variables, such as the patient's background, medical history and current symptoms. The action is the treatment assigned to the patient and a possible response could be the patient's health status after receiving treatment. A more complicated example of a decision making application is a robot learning how to function in its environment. As the robot observes the characteristics of its environment and takes actions, the consequences of those actions (the response) help the robot learn how to interact with its environment and accomplish desired tasks.

Given the goal in decision making, the measure used for comparing alternate policies is called the Value of a policy [53]. The Value of a policy π is the expected sum of responses when following the policy π. Let the distribution of X_t given (H_{t-1}, A_{t-1}) be a fixed distribution with density function f_t. Also let the distribution of R_t given (H_t, A_t) have density function g_t. Then, when actions are chosen according to the policy $\pi = (\pi_1, ..., \pi_T)$ the trajectory $(X_1, A_1, R_1, ..., X_T, A_T, R_T)$ has distribution

$$f_1(x_1)\pi_1(a_1|h_1)g_1(r_1|h_1, a_1)\prod_{t=2}^{T} f_t(x_t|h_{t-1}, a_{t-1})\pi_t(a_t|h_t)g_t(r_t|h_t, a_t). \tag{2.1}$$

If $E_\pi[\ ]$ denotes the expectation with respect to the above distribution, then the

Value of π is then

$$V_\pi = E_\pi \left[\sum_{t=1}^{T} R_t \right] \tag{2.2}$$

The optimal policy, π^*, is then defined as

$$\pi^* = \arg\max_\pi V_\pi = \arg\max_\pi E_\pi \left[\sum_{t=1}^{T} R_t \right] \tag{2.3}$$

Data from sequential decision making applications can come in all shapes and sizes. However, the methods in this thesis were developed for data consisting of a single training set of finite horizon trajectories collected using a known stochastic policy. This policy gives non-zero probability to all possible a_t for any h_t. The methods detailed in this thesis were developed and tested using one time decision problems, however, variable selection for decision making is more critical for problems where the number of time points, $T \geq 0$. Some modifications may be necessary to apply these methods to multiple time point problems; we discuss this in Chapter VI. For the most part we will also assume that the number of trajectories is less than 1000 and the size of the vector h_t is large.

2.1 Characteristics of Decision Making Applications

It is important to understand some basic characteristics that are present in single time point and sequential decision making problems. These characteristics demonstrate how decision making differs from prediction and why finding optimal policies can be difficult.

In a decision making problem, only the response to the chosen action can be observed for each subject. The responses that would have occurred for other actions are unobserved counterfactual outcomes. Without knowing these counterfactual outcomes, the best action for a subject is never really known. We can only infer what

the best action would have been by combining data from many subjects and comparing responses for subjects with similar observations but different actions. In this respect, decision making is different from prediction. Also, the focus of prediction is to find the predictive model with the lowest prediction error. This is different from decision making applications where the focus is choosing the actions that optimize the response. Predictive models may aid in this process but are not the main focus.

In decision making applications with multiple decision time points, $T > 1$, delayed effects may be present. Delayed effects occur when the desirability of a particular action is not manifested in the response immediately following that action. For example, delayed effects occur when the immediate response for a particular action is comparable or worse than alternate actions, but the action leads to better responses in future actions. A real world illustration of delayed effects is the decision to obtain higher education with the result of greater lifetime earning potential. High school graduates who choose to work full time without seeking higher education often initially earn more money than those who choose to go to college. However, those with a college education are likely to earn much more over their lifetime than those who do not. Since the goal in decision making applications is to maximize the sum of responses, it is important to pay attention to delayed effects. Maximizing the response at each time point will not necessarily result in a maximal sum of responses.

Along with delayed effects, in many decision making applications certain actions lead to outcomes that limit or expand the field of possible future actions. In other words, the space of future actions may depend on which actions were taken in the past. Thus, at any given time point in a decision making process, the choice of future optimal decisions may change according to the past actions taken.

2.2 Algorithms for Finding Optimal Policies

Many algorithms exist for finding optimal decisions. This topic has been studied extensively in the field of computer science under the name of reinforcement learning [53, 26] and in the fields of operations research and engineering under the names control theory and dynamic programming [4, 56]. Some research exists in other fields such as statistics under the names adaptive treatment strategies and dynamic treatment regimes [37, 46]. Since the topic of this thesis focuses on variable selection techniques rather than algorithms for finding optimal policies, this section is not meant to be an exhaustive summary of the current available techniques. Rather, it will touch on a few methods that can be used to test variable selection techniques. For more information on algorithms for finding optimal decisions, see [53].

Many methods for finding optimal decisions try to estimate a probability model for the decision making process using maximum likelihood. These models estimate the probability distribution of X_{t+1} and R_{t+1}, given action A_t and history, $H_t = (X_1, A_1, ..., X_{t-1}, A_{t-1}, X_t)$. With an estimated probability model a variety of methods, such as dynamic programming, can be used to solve for the optimal policy [35].

On the other end of the spectrum, there are algorithms that assume little or no knowledge about the probability distribution of the decision making process. Instead, these algorithms search for the best policy among a set of policies, Π, by comparing the estimated values for those policies. This search is most often carried out over a parameterized class of policies, Π, using optimization techniques such as a gradient search [41].

In between the two former groups falls a popular group of reinforcement learn-

ing methods called temporal difference methods [53]. Temporal difference methods work by building models based on temporally successive predictions of the sum of responses. The most popular temporal difference method is called Q-learning and was originally suggested by Watkins [57]. The optimal Q-function at time t is defined as

$$Q_t^*(H_t, A_t) = E_{\pi^*}\left[\sum_{i=t}^{T} R_i \middle| H_t, A_t\right], \tag{2.4}$$

Q_t^* is the expected sum of responses if at time t, action A_t is chosen and the optimal policy is used to choose all subsequent actions. Q_t^* specifies how the future responses depend on (H_t, A_t). Since

$$E\left[\max_{a_t} Q_t^*(H_t, a_t)\right] = E\left[\max_{a_t} E_{\pi^*}\left[\sum_{i=t}^{T} R_i \middle| H_t, a_t\right]\right] = E_{\pi *}\left[\sum_{i=t}^{T} R_i\right],$$

Q_t^* reveals the optimal action at time t for a given H_t, $\pi_t^* = \max_{a_t} Q_t^*(H_t, a_t)$.

The Q-learning algorithm is used to estimate the optimal Q-function. The algorithm is based on the following equations, called the Bellman equations [3]:

$$Q_t^*(H_t, A_t) = E_{\pi^*}\left[\sum_{i=t}^{T} R_i \middle| H_t, A_t\right] = E\left[R_t + \max_{a_{t+1}} Q_{t+1}^*(H_{t+1}, a_{t+1}) \middle| H_t, A_t\right].$$

The Bellman equations show a direct relationship between the optimal Q-function at time t, Q_t^*, and future optimal Q-functions. The Q-learning algorithm takes advantage of this by building a model for Q_t^* based on predictions of the optimal Q-functions in future time points. The Q-learning algorithm is then:

Basic Q-learning Algorithm:

1. Set $Q_{T+1}^*(H_{T+1}, A_{T+1}) := 0$
2. Repeat for each time t = T, ..., 1
   (a) Estimate Q_t^* using some predictive model $\hat{Q}_t$ with (H_t, A_t) as the predictor variables and $R_t + \max_{a_{t+1}} \hat{Q}_{t+1}(H_{t+1}, a_{t+1})$ as the output variable [15]

3. The optimal strategy is any strategy which satisfies

$$\hat{\pi}_t^*(H_t) = arg\max_{a_t} \hat{Q}_t(H_t, a_t), \forall t.$$

The policy $\hat{\pi}^* = (\hat{\pi}_1^*, ..., \hat{\pi}_T^*)$ estimates the optimal policy within the set of policies, Π, parameterized by the chosen predictive model $\hat{Q}_t$ and the set of predictors in (H_t, A_t), $t = 1, ..., T$. While Q-learning does not necessarily provide the global optimal policy, π^* (2.3), it does search for a local optimal policy over the chosen model space. Q-learning, or a close variant of it will be the main learning algorithm used in this thesis to find optimal policies when testing variable selection algorithms.

2.3 Variable Selection for Decision Making

There are multiple reasons why variable selection might be necessary in a decision making application. A few key reasons are, first, inclusion of unimportant variables adds unnecessary noise to the task of learning the optimal policy and inclusion of spurious interactions can lead to bad policies. Thus, careful variable selection could lead to better policies. Second, due to limited resources, many applications can only collect a small number of variables when enacting a policy in a real world setting. For example, when patients go to the doctor for an illness they do not want to be subjected to multiple tests or fill out many questionnaires before the doctor can offer them a treatment. Furthermore, it is often unclear which variables would be most useful and cost-effective to collect. Variable selection techniques could help identify these variables. A third important reason is that the number of possible variables can grow at an alarmingly fast rate with the number of time points. For each time point t, the observation vector may be a large vector. Past observation vectors and actions might also be useful in selecting the action at time t. When we combine these two

groups of variables, along with the possibility of interactions between them, there are numerous possibilities.

It is helpful to categorize variables into one of two different subsets. Some variables in X_t may be highly correlated with the response, R_t. We call these variables *predictive*. *Predictive* variables help reduce the variability in the estimation of Q_t^* (2.4). Occasionally, there will be a small number of variables in X_t that will help pinpoint which actions are optimal. These variables are called *prescriptive* variables [24]. These two categories are not mutually exclusive. We expect most *prescriptive* variables to also be *predictive*, but not vice versa. Both variables are important when trying to find optimal policies, but only *prescriptive* variables are used in the definition of an optimal policy.

For a variable to be *prescriptive*, it must have a qualitative interaction with the best choice of action [42]. In a decision making process with one time point, a variable X_j is said to qualitatively interact with the action, A, if there exists at least two distinct, non-empty sets within the space of X_j for which the optimal action is different. In other words, there exists disjoint, non empty sets $S_1, S_2 \subset space(X_j)$ for which

$$\arg\max_a E[R|X_j = x_{j1}, A = a] \neq \arg\max_a E[R|X_j = x_{j2}, A = a],$$

for all $x_{j1} \in S_1$, and $x_{j2} \in S_2$. Thus, there is a qualitative interaction when the best choice of action changes as the variable X_j changes. In Chapter IV we give more discussion and plots demonstrating qualitative interactions.

There is an abundance of literature discussing qualitative interactions [9, 17, 42, 30, 50, 64]. It is commonly assumed that qualitative interactions are rare in nature and usually have small effects [10, 64]. Much of the biostatistics literature suggests that the search for qualitative interactions should be severely limited and qualitative

interactions that are found should be initially mistrusted [42, 30, 64]. This point of view is fueled by a myriad of papers publishing claims of finding a qualitative interaction through exploratory data analysis, followed by subsequent studies showing contradictory results [20, 64]. We acknowledge this literature and want to discuss why we feel, despite the negative views that have been expressed, that this topic is not a lost cause.

Some of the skepticism concerning the validity of qualitative interactions is due to the way many clinical trials are conducted. The entry criterion for many clinical trials is very strict and the data only represents a small subset of the population for which the treatment may be applied. This is done to minimize the amount of variability in the response that is not directly related to the treatment itself. When strict entry criterion is used, little variation exists in the X matrix and it is often reasonable to assume there are no genuine qualitative interactions over the range of X in the data. However, this does not imply that genuine qualitative interactions do not exist over the range of X for the entire treatable population. For this reason, the methods we present in this thesis are most useful when applied to data that is representative of the entire treatable population (or at least a substantial proportion of the population).

Problems also exist in the way data analyses on clinical trials are sometimes reported. It is tempting in post hoc analyses to comb through the data looking for anything that is significant and interesting partly due to the fact that journals traditionally only publish significant results. Many times this "data fishing" will include looking for significant qualitative interactions. Rarely are the significant values corrected for the number of tests being performed. Thus it is quite reasonable that researchers will find at least one spurious but "significant" qualitative interaction.

This leads to reports in journals claiming the finding of qualitative interactions which do not replicate in subsequent trials. This problem would not occur as often if researchers were more forthcoming about the number of tests they performed and the significance levels they used [2]. However, it is important to note that genuine qualitative interactions with small effect sizes will not be detectable in every data set that is collected, especially those of small size.

Researchers continue to look for qualitative interactions despite the skepticism and warnings about the ills of post hoc analyses. They look for them because one of the underlying goals of clinical research is to find the best treatment for each individual patient. We believe there is a place for post hoc analyses that look for treatment innovations as long as they are done in a more principled fashion. It is our goal to find an approach that assists in finding new qualitative interactions, but is less susceptible to finding spurious results. To address the concern that the methods we create will be equivalent to testing large numbers of interactions with uncorrected significance levels, we include measures taken to control for this in our methods.

It is also important to note that the idea of a qualitative interaction may become more applicable when dealing with multiple time points. The observation vector X_{t+1} is an outcome of the action A_t. It may contain important information about how the subject responds to a certain type of treatment and is more likely to be measured better than the average baseline covariate since it will be less affected by bad memory recall.

We feel it is also important to emphasize that the goal of variable selection techniques is not necessarily to find the 'correct' underlying model. The driving force for the variable selection techniques detailed in this thesis is to find variables which facilitate and improve optimal decision making.

2.3.1 Estimating The Value: A Tool For Assessment

Recall that the goal in decision making problems is to find a policy that optimizes the Value (2.2). One way to decide whether the inclusion of a variable in a model leads to a better policy is to compare the change in Value of the best policy before and after the variable is added. Unfortunately, the underlying distribution of the data (2.1) is usually unknown; so the Value must be estimated. When data is easily collected or simulated, the Value of any particular policy can be simply estimated by enacting the policy and taking the empirical mean of the sum of the responses. However, in clinical studies, data is most often collected under some fixed policy π (e.g. via randomization). Thus estimating the Value of a particular policy π' requires a more sophisticated estimation technique.

A common method for estimating the Value of a policy π', when the policy used to collect the data, π, is known, is an importance sample estimator [47].

$$\hat{V}_{\pi'} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} R_{t,i} W_t(H_{t,i}, A_{t,i}) \tag{2.5}$$

where t denotes time and i denotes trajectory and the weights $W_{t,i}$ are defined as

$$W_t(H_t, A_t) = \prod_{s=1}^{t} \frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}$$

When π' is non-stochastic, this is just an inverse probability weighting estimator [25]. Assuming $\pi_t(a_t|h_t) > 0$ for all t, h_t, and a_t, 2.5 is an unbiased estimator of $V_{\pi'}$.

$$E_\pi\left[R_t\prod_{s=1}^{t}\frac{\pi'_s(A_t|H_t)}{\pi_s(A_s|H_s)}\right]=E_\pi\left[\prod_{s=1}^{t-1}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}E_\pi\left[R_t\frac{\pi'_t(A_t|H_t)}{\pi_t(A_t|H_t)}\middle|H_t\right]\right]=$$

$$E_\pi\left[\prod_{s=1}^{t-1}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}\int\int r_t\frac{\pi'_t(a_t|h_t)}{\pi_t(a_t|h_t)}\pi_t(a_t|h_t)g_t(r_t|h_t,a_t)da_tdr_t\right]=$$

$$E_\pi\left[\prod_{s=1}^{t-1}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}\int\int r_t\pi'_t(a_t|h_t)g_t(r_t|h_t,a_t)da_tdr_t\right]=$$

$$E_\pi\left[\prod_{s=1}^{t-1}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}E_{\pi'}\left[R_t|H_t\right]\right]=$$

$$E_\pi\left[\prod_{s=1}^{t-2}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}E_\pi\left[E_{\pi'}\left[R_t|H_t\right]\frac{\pi'_{t-1}(A_{t-1}|H_{t-1})}{\pi_{t-1}(A_{t-1}|H_{t-1})}\middle|H_{t-1}\right]\right]=$$

$$E_\pi\left[\prod_{s=1}^{t-2}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}\int\int E_{\pi'}[R_t|H_t]\frac{\pi'_{t-1}(A_{t-1}|H_{t-1})}{\pi_{t-1}(A_{t-1}|H_{t-1})}\pi_{t-1}(A_{t-1}|H_{t-1})da_{t-1}\right]=$$

$$E_\pi\left[\prod_{s=1}^{t-2}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}E_{\pi'}\left[E_{\pi'}[R_t|H_t]|H_{t-1}\right]\right]=E_\pi\left[\prod_{s=1}^{t-2}\frac{\pi'_s(A_s|H_s)}{\pi_s(A_s|H_s)}E_{\pi'}\left[R_t|H_{t-1}\right]\right]=$$

$$\ldots=$$

$$E_{\pi'}\left[E_{\pi'}\left[R_t|H_1\right]\right]=E_{\pi'}\left[R_t\right]=$$

Thus,

$$\begin{aligned}E_\pi[\hat{V}_{\pi'}] &= E_\pi\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}R_{t,i}W_t(H_{t,i},A_{t,i})\right]=\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}E_\pi[R_tW_t(H_t,A_t)]\\ &= \frac{1}{n}\sum_{t=1}^{T}\sum_{i=1}^{n}E_{\pi'}[R_t]=\frac{1}{n}\sum_{i=1}^{n}E_{\pi'}\left[\sum_{t=1}^{T}R_t\right]=E_{\pi'}\left[\sum_{t=1}^{T}R_t\right]=V_{\pi'}\end{aligned}$$

When the policy used to collect the data, π, is unknown, this estimator must be adjusted (see [37, 46]). However, for this thesis, we will assume the data comes from a randomized trial, thus the policy used to collect the data is known.

It is important to note that if the assessed policy selects actions that are rare according to the data generating policy, then the number of subjects in the data

whose actions agree with the policy being assessed will be small and consequently the variability of $\hat{V}_{\pi'}$ will be high. Thus, the usefulness of $\hat{V}_{\pi'}$ for variable selection diminishes when either 1) the sample size is small, 2) T is very large, 3) the number of possible actions at each time point is big with respect to the sample size, or 4) the data generating policy is restrictive.

The next chapter reviews the methods currently being used for variable selection in machine learning.

CHAPTER III

Variable Selection

The vast majority of variable selection research focuses on applications dealing with prediction; few techniques, if any, have been developed directly for use in decision making applications. As discussed in the previous chapter, important differences exist between prediction and decision making. Despite these important differences, much insight is gained by studying variable selection techniques developed for prediction. This chapter reviews the current state of variable selection for prediction and highlights which ideas carry through into the decision making setting and which ideas require more development. The chapter begins by defining variable selection and then discusses key issues for successful variable selection. The three types of techniques are presented that have been developed for variable selection in prediction. The chapter concludes with a short discussion section.

Variable selection is the process of selecting the best subset of variables from among a large number of variables. We are given a data set of n observations each consisting of p input variables (features), $X_1, X_2, ..., X_p \in R^n$, and an outcome variable $Y \in R^n$. The goal of prediction is to fit a statistical learning algorithm H on this data such that it can be used to predict new Y' given new observation vectors $X' = (X'_1, ..., X'_p)$. When dealing with prediction, the phrase 'best subset' is

typically interpreted as the smallest subset of variables $X_1^*, ..., X_k^* \in X$ for which the accuracy of a chosen learner H is optimized [28]. In other words, the best subset is the smallest number of variables for which we get the best prediction. This definition is general to allow the terms *accuracy* and *optimize* to be interpreted differently as the characteristics of the data and the learner H vary. The techniques presented in this chapter demonstrate the numerous ways these terms *accuracy* and *optimize* are interpreted to discover the best subset. Note that this definition focuses on improving prediction, not on finding the correct underlying model. Likewise, with decision making, we will focus on maximizing the mean response rather that finding the correct underlying model. We will also refrain from making assumptions about the relative sizes of n to p and p to k, aside from the obvious constraint that $p \geq k$.

The practice of finding the best subset of variables is a difficult task. Many fields of study, such as psychology and medical research, must deal with a high degree of subject to subject variation. This leads to the collection of a myriad of variables in an attempt to cover every possible source of heterogeneity. In some instances, such as genomics and medical image analysis, the number of variables can be much larger than the sample size ($p > n$). The variable selection task grows in complexity with each new variable collected. Trying every possible subset of variables may not be feasible or would often be unwise since it may lead to overfitting [44, 28]. As such, this requires a well-thought-out way to traverse the set of variable subsets to minimize the number of subsets reviewed while still maximizing the probability of finding the optimal subset. This process is approached in many ways. Some techniques create a subset by looking at and selecting variables individually while others focus on adding or eliminating variables from nested subsets of variables that work well together.

Measures that compare subsets must also be determined in order to decipher

which is the optimal subset. Choosing a measure that is too similar to the underlying outcome may also lead to overfitting the data, whereas a measure unrelated to the outcome can lack overall efficiency [28]. Some current techniques take into account a model between X and Y when comparing the subsets; others are model free. Many take into account the chosen learner H, and nearly all use the output Y.

There are multiple reasons why, despite the difficulty, skillful variable selection is advantageous. As with other dimension reduction techniques, it reduces noise and improves the fit of the learning algorithm. Beyond the benefits of dimension reduction in general, unlike feature construction, variable selection techniques approach the task of dimension reduction by directly eliminating a portion of the variables used by a statistical learner. Eliminating variables can reduce the amount of storage required to house the data and the amount of measurement required to complete the learning task. This is a very appealing aspect when the collection of information is costly or time consuming. When fewer variables enter a model, the model is also less complex which often leads to better understanding. It is important to note that although variable selection can be considered a form of model selection, the practice of selecting the optimal set of variables and selecting the optimal model usually must be approached in very different ways. It is not unusual to see these terms used interchangeably in the literature; however, we suggest differentiating them. The variable selection process considers the entire space of variable subsets that could be used given the space of variables collected, whereas the model selection process most often consists of comparing a small number of candidate models selected by the user. The former usually includes a much larger range of possible models and thus very different techniques must be employed.

Variable selection is well developed when dealing with prediction. We follow

the example of machine learning literature by classifying the majority of variable selection techniques for prediction into three types of methods: filters, wrappers, and embedded methods [21, 5, 44]. Keep in mind, as we review these techniques, that the goal of decision making is not to predict well, but rather to optimize the sum of responses by making good decisions. Just because a variable can predict whether the sum of responses will be high or low does not guarantee that the variable will aid in choosing the optimal decisions. These techniques are suited for a different goal. They will, however, give us a framework to use when creating techniques designed explicitly for decision making applications.

3.1 Filter Methods

A filter method for variable selection is any method designed to select variables independently from the chosen statistical learning algorithm. Thus, filter methods may use the outcome Y to select subsets among the different variables $X_1, X_2, ..., X_p$, but not the learning algorithm H. These methods are also known as model free methods and are built to pre-process the data by 'filtering' out variables that are unlikely to be relevant to the chosen outcome or learning task [5]. This lack of dependence on the learning algorithm allows them to be broadly used and makes them less susceptible to the problem of overfitting. Filter methods tend to be faster and less complex than other methods, but are often less accurate. The quickness of these methods seems appealing when the amount of variables to consider is too large for use with more complex integrated methods.

A large proportion of filter methods uses some criterion to rank the importance of each of the variables individually, called variable ranking. These methods then select the top set of ranked features. Some examples of popular ranking criteria are

measures of correlation between the predictor and the outcome or mutual information measures. The algorithm listed below demonstrates variable ranking using Pearson's correlation coefficient, which is equivalent to ranking by the standardized simple linear regression coefficient.

Correlation Variable Ranking Algorithm:

1. For each variable X_i, calculate the Pearson's correlation coefficient with the outcome Y, $r_i = \frac{Cov(X_i,Y)}{\sqrt{Var(X_i)Var(Y)}}$
2. Choose the top k ranked variables X_i based on the ranking $R_i = rank(r_i)$ or choose all variables X_i such that $|r_i| > C$ for some chosen cutoff value $0 < C < 1$.

Since these variable ranking methods look at variables individually, they are usually performed quickly and simply. However, they do not consider whether subsets of variables work well together or whether certain variables are so highly correlated that it is wiser to include only one of them.

Some filter methods use wrapper or embedded methods (see sections 3.2 and 3.3) with very simple models for their learner and then use a more complicated model for the actual learning task [21]. For example, a decision tree could be fit and the variables used by the tree would be considered relevant for use with a more complex learner. This idea is made more flexible by including products of the input variables or higher order transformations for possible selection [21].

The general idea behind filter methods easily transfers over to the decision making setting. Moreover, quick and simple methods that do not account for the chosen learning algorithm are appealing in a setting where learning algorithms are difficult to understand and even more difficult to explain. In this case, however, we should consider more than just the relationship between input and outcome variables. We

also must consider how the relationship between input and outcome variables changes as different actions are taken.

3.2 Wrapper Methods

Wrapper methods for variable selection utilize the chosen learning algorithm when selecting variables. In short, these model based methods try out subsets of variables on the learning algorithm H, then choose the subset which optimizes the predictive power of H. However, directly comparing optimization of the learning algorithm with subsets of variables on a training set alone can lead to overfitting. This problem may be avoided by either using an alternate criteria to judge predictive ability or by splitting the training data into a smaller training set and a validation set for independent testing of the predictive power [28]. The process of training the learning algorithm on different subsets of variables and possibly testing the results on a validation set is computationally expensive. Therefore, strategies that determine the best way to search through the space of variable combinations are crucial with these methods. Many of these search strategies can be classified into three types: 1) forward selection, 2) backward elimination or 3) a combination of forward selection and backward elimination called stepwise procedures. All three types of strategies focus on choosing nested subsets of variables. Forward selection strategies start with an empty set and proceed by adding variables of importance. Backward elimination strategies start with all the possible variables and proceed by eliminating non-beneficial variables.

One well known set of wrapper methods is forward, backward and stepwise regression. These methods, used with H as a linear regression model, add or eliminate features based on p-values, information criterion, or variance measures. Below we list the algorithm for forward selection linear regression using Akaike's Information

Criterion (AIC).

Forward Selection Regression Algorithm:

1. For each variable X_i, compute $AIC_i = nlog(\frac{RSS}{n}) + 2$. Select the variable with the highest AIC_i as the starting variable and set $AIC = AIC_i$ and $p = 2$.

2. For each variable X_i, fit a multiple linear regression with the currently selected variables and X_i, then calculate $AIC_i = nlog(\frac{RSS}{n}) + 2p$

3. If $AIC < AIC_i$, $\forall i$, stop and use currently selected variables

4. Select the variable with the highest AIC_i , set $AIC = AIC_i$ and $p = p+1$, then go back to step 2

The above algorithm constitutes a wrapper method when H is multiple linear regression. Each new subset of variables that is tested is nested within the previous subset of variables selected. In each iteration H is run on the new subsets of variables and the predictive powers are compared by using AIC. AIC measures the strength of fit ($nlog(\frac{RSS}{n})$) versus the complexity of the model ($2p$). It attempts to chose the simplest model that fits the data well. For more information on AIC see [1].

Wrapper methods are probably the easiest techniques to adjust for the setting of decision making. They essentially treat the learning process as a black box [28]; thus it wouldn't matter whether the goal of the model was optimal prediction or optimal decision making as long as we had a good measure to assess the fit of the model. The big drawback to these methods, as suggested earlier in this section, is their computational cost. This becomes even more of an issue with decision making algorithms. Learning algorithms for decision making often require multiple predictive models in order to estimate the optimal policy. In addition, the Value of the policy must be estimated for model assessment.

3.3 Embedded Methods

Variable selection methods built into the learning algorithm H, are called embedded methods. These methods simultaneously choose the optimal subset of variables while fitting the learning model. They are designed to work with a specific type of learning algorithm and are not as generalizable as wrapper or filter methods. However, these methods are typically much less computationally expensive than wrapper methods, and tend to have better predictive ability than filter methods.

One area of embedded methods is the set of learning algorithms that optimize a penalized loss function or a penalized likelihood function using a penalty function that leads to sparse representations, such as the L_0 or L_1 norm [6, 16, 54]. A well known example from this set is the Lasso model [54]. This model uses the sum of squared errors of a linear model on the parameters as its loss function and the L_1 norm of the parameter coefficients as its penalty function. See equation (3.1) below for its numerical formulation.

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{n} (Y - \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_i| \tag{3.1}$$

The L_1 norm penalty causes the coefficient vector, $\hat{\beta}$, to be sparse. In other words, many $\hat{\beta}_i$ are exactly zero, so that the predictive model, $\hat{Y_i'} = \hat{\beta}_0 + \sum_{j=1}^{p} X_{ij}'\hat{\beta}_j$, will only depend on a portion of the predictors. The variable λ in the formulation above is called a tuning parameter. This tuning parameter determines the sparseness of the $\hat{\beta}$ vector. Equation (3.1) can be solved using quadratic programming. For more information on the Lasso see [23].

Another example of a popular embedded method is classification and regression trees [7]. Classification and regression tree algorithms recursively partition the obser-

vation space to create subsets of similar observations based on important predictors and responses. At each node the optimization criteria chooses the best variable to partition the data. Typically only a subset of the candidate variables will be chosen as partitioning variables. Only the variables that are used in a partition are needed and the remaining variables can be left behind. Thus, tree algorithms perform variable selection while fitting a model.

Developing embedded methods designed specifically for decision making applications will probably take more thought and consideration than the previous two types of methods. Since most algorithms for calculating optimal decisions include steps that estimate a predictive model, embedded methods for prediction can be incorporated at these steps. More research is needed, though, to determine how this might effect other steps in the algorithm and the overall performance of the model.

3.4 Hybrid Methods

Each of the three prior methods work best in different situations. Filter methods tend to be quick and easy, working best when the number of variables is extremely large. Wrapper methods are employable with most learning algorithms when the number of variables is not excessive. Embedded methods are an elegant alternative when the other two types are unappealing. Research exists on combining techniques to gain the advantages of both [11, 60]. For example, we might use a filter method first to trim the number of variables, then use a wrapper method on the trimmed set to maximize predictive power with minimal computational expense. These combined methods are often referred to as 'hybrid' methods.

3.5 Bayesian Variable Selection

Bayesian variable selection methods also exist. While we will not discuss them thoroughly in this thesis, we want to make the reader aware of their existence. Most of these techniques can be considered a type of embedded or wrapper method, and many of the non-Bayesian embedded and wrapper methods can be equivalently expressed using a Bayesian formulation. As suggested in the section on filter methods, one could also use Bayesian methods as a pre-processing step to choose the variables and then separately fit a different model on the chosen variables. While Bayesian methods are potentially useful in variable selection for decision making, we leave research in this area for future work. For more information on Bayesian variable selection techniques we recommend [8, 12, 18, 19].

3.6 Discussion

In this chapter we discussed variable selection techniques for prediction. Currently, the techniques used for variable selection in sequential decision making are primarily guided by expert opinion. However, some areas lack sufficient domain knowledge and expertise to determine which variables are best. There are also a few cases where predictive variable selection techniques were used, such as Lasso and decision trees [15, 34]. In medical decision making applications such as clinical trials, predictive methods are commonly used for variable selection. There are a few qualitative interaction tests that can be used to test a small number of expert determined pre-specified interactions [50, 17, 39, 51, 38, 29, 61, 62, 45]. These tests are too conservative to be used on a large set of interactions when controlling the error rate for multiple testing [17, 43, 62]. Also, many of the tests were designed for testing only qualitative interactions between categorical variables and the treatment action.

The topic of variable selection designed for sequential decision making has received little attention. We believe all three types of methods discussed in this chapter show good potential for use in decision making problems with adjustments made to the final goal. The next chapter presents two hybrid type methods that are similar to methods for prediction, but are designed to find variables that are important for decision making.

CHAPTER IV

Variable Selection for Qualitative Interactions

This chapter describes two new methods to select variables useful for decision making. Applications that deal with decision making occur in many different fields such as computer science, engineering, economics and medicine. In medicine, deciding when a patient needs treatment and which treatment is best are critical decisions. Clinical trials, particularly those involving heterogenous patients, collect a large amount of potentially useful information that can aid in making these decisions. However, in clinical practice much of this information is expensive, time-consuming, and/or burdensome to collect. Thus variable selection is needed to help inform the clinicians which variables are most important.

Variable selection techniques have been developed to enhance prediction, but their use in decision making has not been well tested. In our research we have found these techniques often miss or down play the importance of certain interaction variables that are key to making decisions. The variable selection techniques we propose focus on finding these important interactions.

This work is motivated in part by the the Nefazodone CBASP trial data. The Nefazodone CBASP trial [27] was a randomized controlled trial conducted to compare the efficacy of three alternate treatments for patients with chronic depression. The

study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. Analysis of the trial data showed the combination treatment to be superior to the two singleton treatments overall. We wanted know whether this relationship held true for all subsets of patients, and if not, to discover which patient characteristics help to determine the optimal depression treatment for an individual patient.

The remainder of this chapter is organized as follows: Sections 2 and 3 give background material on optimal decision making and discuss what makes a variable important for decision making. Section 4 provides two ranking techniques designed to find variables useful for decision making followed by an algorithm for using these techniques. Section 5 presents some simulation experiments, and Section 6 illustrates the methods using data from the Nefazodone CBASP study. Concluding remarks are given in Section 7.

4.1 Optimal Decision Making

We consider variable selection in the simplest decision making setting in which one must decide between two actions. The idea is to use observations about a subject $X = (X_1, X_2, ..., X_p)$, to choose a treatment action A. Following the action a response occurs. The response, R, gives us some indication of the desirability of the chosen action. The goal is to choose actions that maximize the response. A policy, π, is a stochastic or deterministic decision rule mapping the space of observations, X, to the space of the action, A. In other words, π defines the probability for choosing action $A = a$ given the observations $X = x$. So the goal can be restated as finding a policy π^*, that maximizes the response.

A simple example of a decision making problem is a clinical trial to test two alternative drug treatments. The observation vector, X, would consist of baseline variables, such as the patient's background, medical history and current symptoms. The action would be the treatment assigned to the patient and the response could be the patient's condition or symptoms after receiving treatment. The goal is to determine which treatment is optimal for any given future patient, using the data obtained in the trial.

Alternate policies can be compared via the expected mean response, called the Value of a policy [53]. Let the distribution of X be a fixed distribution f, and let the distribution of R given (X, A) be a fixed distribution g. Then when actions are chosen according to a policy π, the trajectory (X, A, R) has distribution

$$f(x)\pi(a|x)g(r|x,a), \tag{4.1}$$

If $E_\pi[\ ]$ denotes the expectation over the above distribution, then the Value of π is

$$V_\pi = E_\pi\left[R\right]$$

The optimal policy, π^*, is then defined as

$$\pi^* = \arg\max_\pi V_\pi = \arg\max_\pi E_\pi\left[R\right],$$

or equivalently

$$\pi^*(x) = \arg\max_a E\left[R|\, X = x, A = a\right].$$

If we knew the multivariate distribution of (X,A,R), the best treatment for future use could be found by calculating $E\left[R|\, X = x, A = a\right]$ for every possible (x, a) combination and then selecting the action leading to the highest conditional expectation of R for each x. In practice, however, we do not know this distribution. So we must use data to estimate the optimal future treatment. We do this by first estimating

$E[R|\, X = x, A = a]$ for each (x, a) using a predictive model and learner, such as a multiple linear regression. We then use the estimated regression function to 'estimate' the best future treatment for each x. For example, if we used the data to estimate $E[R|\, X = x, A = a]$ by

$$\hat{E}[R|\, X = x, A = a] = \hat{\beta}_0 + x\hat{\beta}_1 + a\hat{\beta}_2 + xa\hat{\beta}_3,$$

for $a \in \{0, 1\}$, our estimated optimal future treatment actions would be

$$\hat{\pi}^*(x) = I(\hat{\beta}_2 + x\hat{\beta}_3 > 0).$$

4.2 Variable Selection

There are multiple reasons why variable selection might be necessary in a decision making application. One reason is that finding the optimal policy becomes more difficult as the number of spurious variables included in the model increases. Thus, careful variable selection could lead to better policies. Also, due to limited resources, only a small number of variables may be possible to collect when enacting a policy in a real world setting. Researchers are often unsure which variables would be most important to collect. Variable selection techniques could help identify these variables. In addition, policies with fewer variables are often easier to understand, so variable selection can improve interpretability.

Currently, variable selection for decision making in many fields is predominantly guided by expert opinion. Expert opinion can be a good starting place when there is sufficient domain knowledge and expertise. Some predictive variable selection techniques, such as Lasso [54], have been suggested [34]. In clinical trials, a combination of predictive variable selection techniques and statistical testing of a small number of interaction variables suggested by expert opinion are most commonly used

[38, 29, 45]. Little research has been carried out to evaluate these techniques in decision making or suggest how they might be improved.

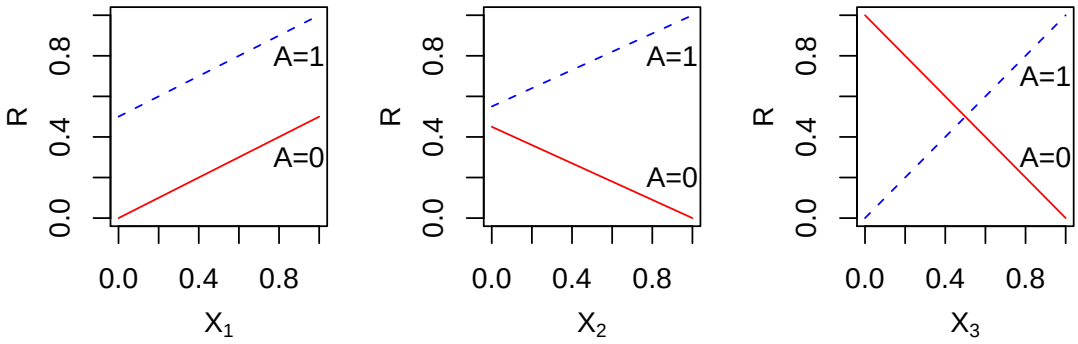
When selecting variables for decision making, a distinction should be made between variables that are included merely to facilitate estimation as opposed to variables involved in the decision rules. *Predictive* variables are variables used to reduce the variability and increase the accuracy of the estimator. Variables that help prescribe the optimal action for a given patient are *prescriptive* variables [24]. For optimal estimation results, it is best to select both types of variables. However, only *prescriptive* variables need to be collected when implementing the policy.

For a variable to be *prescriptive*, it must have a qualitative interaction with the action [42]. A variable X_j is said to qualitatively interact with the action, A, if there exists at least two distinct, non-empty sets, $S_1, S_2 \subset space(X_j)$ for which

$$\arg\max_a E[R|X_j = x_{j1}, A = a] \neq \arg\max_a E[R|X_j = x_{j2}, A = a],$$

for all $x_{j1} \in S_1$, and $x_{j2} \in S_2$. These variables are useful for decision making because they help decipher which action is optimal for each individual patient.

To illustrate this idea, see the plots in Figure 5.1. These plots depict different possible relationships between the conditional mean of R, A and a particular X_j, when averaging over all other X_i, $i \neq j$. Figure 5.1(a), shows a variable, X_1, which does not interact with the action. Figure 5.1(b) shows a variable, X_2, that interacts with the action, A, but does not qualitatively interact with the action. In both plots, the optimal action is $A = 1$. Knowledge of X_1 or X_2 is useful for predicting the response for a given action, but should not affect which action should be chosen. Figure 5.1(c), shows a variable, X_3, which qualitatively interacts with the action. We can see that the optimal action in this plot is $A = 0$, when $X_3 \leq .5$ and $A = 1$ when $X_3 > .5$. Knowledge of X_3 impacts the best choice of the action and likewise



(a) No interaction (b) Non-qualitative interaction (c) Qualitative Interaction

Figure 4.1: Plots demonstrating qualitative and non-qualitative interactions

the response, thus it is important for decision making.

The degree to which a prescriptive variable is useful depends on two factors:

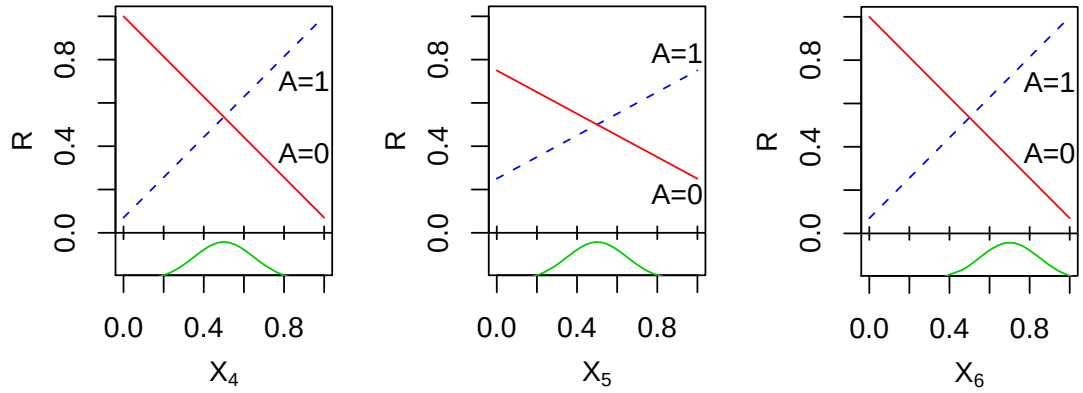
1. *Interaction*: the magnitude of the interaction between the variable and the action. For an action with two possible values, $A \in \{0, 1\}$, this is the degree to which the following quantity varies as x varies

$$E\left[R|X = x, A = 1\right] - E\left[R|X = x, A = 0\right] \tag{4.2}$$

2. *Proportion*: the proportion of patients whose optimal choice of action changes given a knowledge of the variable. If $a^* = \arg\max_a E\left[R|A = a\right]$, this is the proportion of patients for which the following holds:

$$\arg\max_a E\left[R|X = x, A = a\right] \neq a^* \tag{4.3}$$

Consider the plots in Figure 4.2. Figure 4.2(a) shows the relationship between the conditional mean of R, A, and a variable X_4, with an underlying plot giving the distribution of X_4. Figures 4.2(b), 4.2(c) are similar to Figure 4.2(a), but for variables X_5 and X_6. Notice that X_4 and X_5 have the same distribution. However, the interaction between X_4 and A is much stronger than the interaction between X_5



(a) Large interaction large proportion

(b) Small interaction, large proportion

(c) Large interaction, small proportion

Figure 4.2: Plots demonstrating usefulness factors of qualitative interactions

and A. Therefore, the effect of choosing the optimal action is much greater given X_4 than it is given X_5. Now notice that X_4 and X_6 have the same relationship with the conditional mean of R and A but are distributed differently. The distribution of X_4 is centered at the point of intersection, so half of the subjects would do better choosing $A = 0$ over $A = 1$. Whereas, the proportion of patients benefiting from choosing $A = 0$ is much smaller with X_6. Thus X_4 would be more useful in decision making than X_5 or X_6.

Since both of these factors also affect the predictive ability of a qualitative interaction, it may not be readily apparent why current variable selection techniques designed for prediction are not well equipped to find these prescriptive variables. One reason current variable selection methods aimed at prediction may have problems detecting prescriptive variables may be due to the way prescriptive variables occur in nature. In real world applications, individual variables, rather than interactions between variables, tend to explain most of the variation in the outcome and thus are most important for good prediction. This individual effect a variable has on the response is often referred to as the 'main effect' of the variable and most

variable selection techniques are good at finding main effects. Furthermore, while non-qualitative treatment-covariate interactions do occur quite frequently in real world applications, it is commonly assumed that qualitative interactions are rare in nature [10, 64].

There is an abundance of literature discussing qualitative interactions (e.g. Byar and Corle, 1977; Gail and Simon, 1985; Peto, 1982; Lagakos, 2001; Shuster and Van Eys, 1983; Yusuf et al., 1991; Senn, 2001). Much of the statistical literature suggests that the search for qualitative interactions should be severely limited and qualitative interactions that are found should be initially mistrusted [42, 30, 64, 48]. This point of view is fueled by a myriad of papers publishing claims of finding a qualitative interaction during exploratory data analysis of a controlled trial followed by subsequent studies in which the interaction did not replicate [20, 64].

Skepticism concerning the validity of qualitative interactions is partially due to the way many clinical trials are conducted. Entry criterion are restrictive for many clinical trials. This results in data with minimal variation in the X variables, representing only a small subset of the treatable population. In this case it is often reasonable to assume there are no genuine qualitative interactions within the range of the data. However, this does not imply that genuine qualitative interactions do not exist over the range of X for the entire treatable population. For this reason, the methods we present are most useful when applied to data representative of the entire treatable population (or at least a substantial proportion of the population).

Skepticism also exists due to the way analyses of clinical trials are reported. Since journals traditionally publish only significant results, it is tempting to comb through the data in post hoc analysis looking for anything that is significant and interesting. Many times this "data fishing" includes looking for significant qualitative interac-

tions. When significance levels are not corrected for the number of tests performed, researchers can often find at least one significant qualitative interaction spuriously. This problem would not occur as much if researchers were more forthcoming to journals about the number of tests they performed and the significance levels they used [2]. However, it is important to note that genuine qualitative interactions with small effect sizes will be undetectable in some data sets, especially those of small sample size. Despite this skepticism, medical researchers continue to look for qualitative interactions. They look for them because it is an underlying goal of clinical research to find the best treatment for each individual patient

Our goal is to develop approaches that assist in finding qualitative interactions, but are less susceptible to finding spurious results. To address the concern that the proposed methods are equivalent to testing large numbers of interactions with uncorrected significance levels, we thoroughly test them on simulated data generated without qualitative interactions. Beyond this, we feel it is also important to emphasize that the goal of these methods is not to find the 'correct' underlying model. Rather, the driving force for this variable selection is to facilitate and improve decision making by reducing the number of variables that need to be considered when constructing a policy.

4.3 Qualitative Interaction Ranking

As we discussed in the previous section, variable selection in decision making should focus on variables that qualitatively interact with the action. In this section we will present two variable ranking techniques that rank the variables in X based on their potential for a qualitative interaction with the action variable. We conclude the section with a proposed complete algorithm for variable selection in a decision

making application.

The first variable ranking method is based upon the two usefulness factors for a qualitative variable discussed in the previous section (see quantities (4.2) and (4.3)). Assume we have a data set of n subjects, with p baseline observations taken on each subject, making up the $n \times p$ observation matrix X. Also assume that in the data the action, $A = \{0, 1\}$ is randomized. The response is denoted by R. Consider the evaluation of the jth variable, X_j (the jth column of X). Then given an estimator of $E[R|X_j = x_j, A = a]$ say $\hat{E}[R|X_j = x_j, A = a]$, define the following quantities for $j = 1, ..., p$:

$$D_j = \left(\max_{1 \le i \le n} \left(\hat{E}[R|X_j = x_{ij}, A = a^*] - \hat{E}[R|X_j = x_{ij}, A \neq a^*] \right) - \min_{1 \le i \le n} \left(\hat{E}[R|X_j = x_{ij}, A = a^*] - \hat{E}[R|X_j = x_{ij}, A \neq a^*] \right) \right) \tag{4.4}$$

and

$$P_j = \frac{1}{n} \sum_{i=1}^{n} 1\{\arg\max_a \hat{E}[R|X_j = x_{ij}, A = a] \neq a^*\} \tag{4.5}$$

where $1\{.\}$ is 1 if '.' is true and 0 otherwise and $a^* = \arg\max_a \hat{E}[R|A = a]$ is the overall optimal action.

D_j is a measure of the magnitude of the interaction. P_j is a measure of the proportion of subjects affected by a change in the optimal choice of action due to the inclusion of an interaction involving X_j. These two quantities can be combined to make a score, U_j, for ranking the variables:

$$U_j = \left(\frac{D_j - \min_{1 \le k \le p} D_k}{\max_{1 \le k \le p} D_k - \min_{1 \le k \le p} D_k} \right) \left(\frac{P_j - \min_{1 \le k \le p} P_k}{\max_{1 \le k \le p} P_k - \min_{1 \le k \le p} P_k} \right) \tag{4.6}$$

The first term in parentheses provides the relative (as compared to the other variables in X) magnitude of X_j's interaction with the action; the second term in parentheses

provides the relative proportion of affected subjects in X_j used to select the action. U_j is a product because we want to select X_j only if both D_j and P_j are relatively large. The first variable ranking procedure will rank variables in terms of their U_j.

The second ranking procedure looks directly at the expected increase in the estimated optimal Value due to the knowledge of the variable X_j. It estimates the quantity described by [40] as the value of information. Define the score S_j as

$$S_j = \sum_{i=1}^{n} \left[\max_a \hat{E}\left[R|X_j = x_{ij}, A = a\right] - \hat{E}\left[R|X_j = x_{ij}, A = a^*\right] \right] \tag{4.7}$$

Both of these scores, U and S can be used to rank the variables. They have been defined generically to allow different models for $E[R|X, A]$. In the numerical section that follows, we use a linear model to estimate the conditional expectation and obtain $\hat{E}$.

Although not explicitly shown in the notation, predictive variables may also be used in the estimation of the conditional expectation. When testing for the interaction between X_j and A, researchers often prefer to maintain a hierarchical ordering [59] and thus the main effect of the variable X_j and the main effect of the action should be included. This helps to avoid finding spurious interactions that may appear because the main effect is important but is not included in the estimation. It is also wise to include other important main effects of the variables in X on R to help reduce variability in the estimation.

4.3.1 Variable Selection Algorithm

The following is an overview of an algorithm for variable selection.

1. **Select Important Predictors:** Select important predictive variables of R among $(X, A * X)$ using a Lasso with the penalty parameter chosen by Bayesian

Information Criterion (BIC)

2. **Rank Interactions Individually:** Rank the variables in X using either U or S. Use the main effect variables selected in step 1 to help decrease the variability in the estimator $\hat{E}$. Select the top H variables in rank, where $H =$ the number of variables having non-zero U or S scores.

3. **Create Nested Subsets of Chosen Predictive and Prescriptive Variables:**

   (a) Collect the following K variables:

      i. The predictive variables chosen in step 1 and

      ii. The main effects of the top H ranked variables in step 2 and

      iii. The interactions between A and the top H ranked variables in step 2

   (b) Run a weighted Lasso using a weighting scheme that satisfies the following properties

      i. All main effect variables and all interaction variables chosen in step 1 only are given a weight $w = 1$

      ii. All interaction variables chosen in step 2 are given a weight $0 < w \leq 1$ which is a non-increasing function of the U or S score

   (c) Create K nested subsets based on the order of entry of the K variables in the weighted Lasso in the previous step

4. **Select Subset Using Adjusted Gain in Value Criterion:**

   (a) For each subset $k = 1, ..., K$, estimate the maximal Value, e.g.

      i. Use the subset to estimate $\hat{E}$

      ii. Estimate the optimal policy, $\hat{\pi}_k^*(x) = \arg\max_a \hat{E}\left[R|X = x, A = a\right]$

iii. Estimate the Value of $\hat{\pi}_k^*$ by:

$$\hat{V}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{E}[R|X = x_i, A = \hat{\pi}_k^*(x_i)]$$

(b) Select the subset, k^*, that has the highest Adjusted Gain in Value (AGV) criterion:

$$AGV_k = \frac{\hat{V}_k - \hat{V}_0}{\hat{V}_m - \hat{V}_0}\left(\frac{m}{k}\right)$$

where $m = \arg\max_k \hat{V}_k$ and $\hat{V}_0$ is the estimated Value of the policy $\hat{\pi}_0^* = \arg\max_a \hat{E}\left[R|A = a\right]$

In step 1 we use Lasso to find the variables among $(X, A * X)$ that are important predictors of R. We chose Bayesian Information Criterion to select the penalty parameter (Zou, Hastie and Tibshirani, 2007) because of its conservative nature to ensure only strong predictors enter the model. Predictive variables are important for reducing variability in the estimations. However, predictive variables are only part of the puzzle, so we add to step 1 a few more steps to help our algorithm select both prescriptive and predictive variables. In step 2 we look for qualitative interactions individually using an approach which rates each variable in X based on its potential for a qualitative interaction with the action. We look at each of the interaction variables individually to avoid problems with collinearity. In steps 3 and 4 we seek to further refine the set of variables collected in steps 1 and 2. In step 3 we seek a quick way to navigate through the space of all possible combinations of the variables collected in steps 1 and 2. Thus we chose to create nested subsets from the variables based on order of selection in a weighted Lasso. This ordering by the weighted Lasso gives us a joint ranking of all the variables selected in steps 1 and 2. We use the weighting scheme in the weighted Lasso to balance the importance of both predictive and prescriptive variables in the decision making process. Since Lasso

favors variables that are predictive we offset this by down-weighting the prescriptive variables. In step 4 we select between the different subsets using the AGV criterion, a criterion that trades off between the complexity and the observed Value of each of the models.

The AGV criterion selects the subset of variables with the maximum proportion of increase in Value per variable. It is similar in idea to the adjusted R^2 value. The model with $m = \arg\max_k \hat{V}_k$ variables is akin to a saturated model, because the addition of more variables does not improve the Value of the model. Thus the denominator is the observed maximum gain in value, among the different variable subsets, divided by m, an estimate of the degrees of freedom used to achieve that gain in Value. The numerator then measures the gain in Value of the intermediate model, the model with k variables, divided by k, the estimated degrees of freedom needed to achieve that gain in Value.

An alternate way to look at the AGV criterion is that the quotient $(\hat{V}_k - \hat{V}_0)/(\hat{V}_m - \hat{V}_0)$ compares the gain in value for the current subset of variables against the maximum gain in value over all the subsets of variables. Ideally this term stays fairly stationary whenever a main effect variable is added to the model and increases when a qualitative interaction is added to the model. Thus this quotient is expected to be approximately monotone increasing with k. The quotient, m/k, acts as a penalty on the inclusion of variables that do not substantially increase the Value. We include main effect variables in the counts m and k because each main effect variable that is included decreases the degrees of freedom. Also, the inclusion of main effects in the counts quickly deflates the quotient as k increases leading to a less severe penalty on larger models. This is helpful since there is often many more useful predictive variables than prescriptive variables.

In the next section we test this algorithm on simulated data. We reference the algorithm as Method U or Method S depending on the scoring function U or S that was used in step 2. For the weighting scheme in step 3(b) we tried multiple different schemes (inverse, exponential, etc.). In practice the weighting scheme that worked best is listed below:

1. All predictive variables are given a weight $w = 1$
2. All prescriptive variables are given a weight $w = 1 - \frac{U}{max(U)+\epsilon}$ or $w = 1 - \frac{S}{max(S)+\epsilon}$ respectively

The ϵ term in the weight is needed to ensure $w \neq 0$. So ϵ can be thought of as a stabilizing factor, but it can also be thought of as the balancing factor between prescriptive and predictive variables (i.e. large ϵ favors predictive variables, small ϵ favors prescriptive variables). In experimentation we found $\epsilon = H/n$ to be a good value.

4.4 Simulations

To test the performance of the new techniques, we ran them on realistically designed simulation data and compared the results to using Lasso [54]. Lasso was used to select from the set of main effects of X, and the interactions between A and each variable in X. The main effect of A was not subject to selection, that is the coefficient of A was unconstrained by the L_1 penalty function. We tested 2 different methods for choosing the penalty parameter. The first method we used was the Bayesian Information Criterion (BIC) as defined in Zou, Hastie and Tibshirani [67]. We reference this method as BIC Lasso. Zou et al. [67] recommend this method when using Lasso primarily for variable selection. The second method we used for choosing the penalty parameter was 5-fold cross-validation on the prediction error of

the Lasso model [54]. This is a standard method for choosing the penalty parameter and we reference this method as CV Lasso. Note that our method uses Lasso as well, but only to select predictive variables in step 1 and to order variables in step 3(a).

To generate realistic simulation data, we randomly selected rows, with replacement from X, the observation matrix from the Nefazodone CBASP trial data. We generated new actions, A, and new responses, R, that covered a wide variety of models. We report results for the following generative models:

1. Main effects of X only, no treatment effect and no interactions with treatment
2. Main effects of X, moderate treatment effect and no interactions with treatment
3. Main effects of X, moderate treatment effect, multiple medium to small non-qualitative interactions with treatment, no qualitative interaction with treatment
4. Main effects of X, small treatment effect, small qualitative interaction with a binary variable, no non-qualitative interactions
5. Main effects of X, small treatment effect, small qualitative interaction with a continuous variable, no non-qualitative interactions
6. Main effects of X, small treatment effect, multiple moderate to small non-qualitative interactions with treatment, small to moderate qualitative interaction with a binary variable and treatment
7. Main effects of X, small treatment effect, multiple small non-qualitative interactions with treatment, small qualitative interaction with a continuous variable and treatment

For each generative model, we used main effect coefficients for the variables X, estimated in an analysis of the real data set. In generative models 3-7 we randomly selected variables from the Nefazodone CBASP data for each treatment covariate interaction and used these same variables for each repetition. The treatment, qualitative interaction and non-qualitative interaction coefficients were set using a variant of Cohen's D effect size measure [10] shown below:

$$D = \frac{\beta\sqrt{Var(R)}}{\sqrt{Var(X_j)}} \tag{4.8}$$

We altered this formula by replacing the marginal variance, $Var(R)$, with the conditional variance of the response $Var(R|X, A)$. However, we maintained the definitions of 'small' and 'moderate' effect sizes suggested by Cohen [10] as $D = 0.2$ and $D = 0.5$ respectively. Thus the effects are slightly smaller than the traditional definition.

For each generative model, we ran CV Lasso, BIC Lasso, Method U and Method S to see which interaction variables were selected by each method. We repeated this 1000 times and recorded the percentage of time each variable was selected for each method and the sign of the coefficient of each interaction selected.

For each repetition, we also calculated the following statistic for each method

$$T = \frac{V_{\hat{\pi}^*} - V_{\pi}}{V_{\pi^*} - V_{\pi}}$$

where V_{π^*} is the Value of the true optimal policy, π^*, V_{π} is the Value of an 'agnostic' policy π which gives equal probability to each action and $V_{\hat{\pi}^*}$ is the Value of the estimated optimal policy given the selected variables. We estimated the policy $\hat{\pi}^*$ by first fitting a linear model of the selected variables on the response using the training set and then optimizing the fitted model with respect to the action.

The statistic T gives the percentage of gain in Value when using the estimated optimal policy as opposed to an 'agnostic' policy relative to the percentage of gain

Table 4.1: Simulation results: BL stands for BIC Lasso, U for method U and S for method S. The first two columns summarize the difference in percentage statistics T between BIC Lasso and the two new methods; values denoted with a * are significantly different from zero using a two-sided t-test with $\alpha = .05$. Note: model 1 has no treatment effect or interactions with treatment, thus all policies return the same Value. The next three columns give the average number of spurious interactions selected by the three methods over the 1000 repetitions. The last three columns give the selection percentage of the qualitative interaction (when one existed) for each method.

Generative	**Ave**		**Ave # of Spur. Interact.**			**Selection Percentage**		
Model	$T_U - T_{BL}$	$T_S - T_{BL}$	BL	U	S	BL	U	S
1	NA	NA	0.04	1.9	1.3	-	NA	-
2	-0.027^*	-0.025^*	0.03	0.6	0.5	-	NA	-
3	0.000	0.000	0.4	0.6	0.5	-	NA	-
4	0.212^*	0.322^*	0.1	1.7	1.0	6	24	27
5	0.280^*	0.226^*	0.1	1.2	1.2	6	35	27
6	0.219^*	0.387^*	0.1	1.0	0.3	25	53	74
7	0.128^*	0.103^*	0.1	0.9	0.8	12	60	49

in Value when using the true optimal policy as opposed to an agnostic policy. We compared the new methods with both of the Lasso competitors by looking at the difference in their T statistics. The results are listed in Tables 4.1 and 4.2. Differences denoted with a * are significantly different from zero using a two sided t-test with $\alpha = .05$. Note that since generative model 1 has no treatment effect and no interactions with treatment, all policies will have the same Value resulting in an undefined T statistic. The tables also list the average number of spurious interactions selected by each method and the selection percentage of the qualitative interaction (if one existed) over the 1000 repetitions.

Looking over Table 4.1 we see that BIC Lasso tends to include a slightly smaller number of spurious interactions, as expected, due to its conservative nature [67]. It's conservative nature is also a bonus in terms of the average Value in the rare situation when no interactions exist in the generative model (generative model 2). However, the use of this method results in a dramatic loss in the average Value when a qualitative interaction does exist because the qualitative interaction is often left

Table 4.2: Simulation results: CL stands for Cross-validated Lasso, U for method U and S for method S. The first two columns summarize the difference in percentage statistics T between CV Lasso and the two new methods; values denoted with a * are significantly different from zero using a two-sided t-test with $\alpha = .05$. Note: model 1 has no treatment effect or interactions with treatment, thus all policies return the same Value. The next three columns give the average number of spurious interactions selected by the three methods over the 1000 repetitions. The last three columns give the selection percentage of the qualitative interaction (when one existed) for each method.

Generative	**Ave**		**Ave # of Spur. Interact.**			**Ave Selection Percentage**		
Model	$T_U - T_{CL}$	$T_S - T_{CL}$	CL	U	S	CL	U	S
1	NA	NA	4.9	1.9	1.3	-	NA	-
2	0.021*	0.022*	4.8	0.6	0.5	-	NA	-
3	0.009*	0.009*	8.4	0.6	0.5	-	NA	-
4	0.031*	0.140*	5.2	1.7	1.0	40	24	27
5	0.113*	0.059*	4.6	1.2	1.2	37	35	27
6	0.095*	0.263*	5.0	1.0	0.3	69	53	74
7	0.097*	0.072*	5.6	0.9	0.8	57	60	49

out.

Table 4.2 shows that while CV Lasso is good at selecting the qualitative interaction, it tends to include several more spurious interactions than the new methods. This leads to a significant loss in the average Value due to policies with bad decisions based on the spurious variables selected when using this method.

Overall, we found that the two new methods perform well. While, the competing Lasso methods each have their appeals in terms of selection, both are less appealing than the new methods when considering the average Value returns and the purpose of the variable selection.

4.5 Application: Nefazodone CBASP Trial

To apply this method to a real data set we suggest augmenting this algorithm in two ways. First we use bootstrap sampling [14] of the original data to give a measure of reliability on the results. That is, take 1000 bootstrap samples of the data, run the algorithm and record the interaction variables that are selected along with the sign

of the interaction coefficient for each bootstrap sample. This will give a percentage of time each interaction variable is selected by the method. Define the adjusted selection percentage to be the absolute value of the number of times an interaction is selected with a positive coefficient minus the number of times an interaction is selected with a negative coefficient. This adjustment eliminates variables that, across the bootstrap samples, do not consistently interact in one direction with the action.

Second we construct a threshold to determine which interaction variables to include in the final model. The threshold estimates the selection percentages we would expect to see if the data contained no interactions. To compute the threshold, we first remove the interaction effects within the data by randomly reassigning the observed values for the interaction variables to different subjects. In other words, permute the X values of the $X*A$ interactions in the $(X, A, X*A)$ model matrix. After obtaining 100 permuted data sets, on each permuted data set we run the same analysis of taking 1000 bootstrap samples, running the algorithm and recording the selection percentage of each interaction variable over the 1000 bootstrap samples. We then calculate the maximum adjusted selection percentage over the p interaction variables for each permuted data set. The threshold is then set to be the $(1-\alpha)th$ percentile over the 100 maximum selection percentages. We found in simulations that the threshold effectively controlled the family-wise error rate to be approximately α giving us in any given experiment $(1-\alpha)\%$ confidence that a variable with a selection percentage above this threshold interacts with the action.

This augmentation by bootstrap resampling and thresholding helps to stabilize the results and it is possible to apply it to other variable selection algorithms, not just the new methods suggested in this paper. In simulations we found the bootstrap resampling and thresholding also effectively controlled the family-wise error for BIC

Lasso. The bootstrap resampling can also be done with CV Lasso, however, this threshold ended up far too conservative to control the family-wise error rate for the CV Lasso.

To demonstrate these new methods along with this augmentation we applied them to a real data set dealing with depression. As introduced previously, the Nefazodone CBASP trial [27] was conducted to compare the efficacy of three alternate treatments for patients with chronic depression. We applied the methods to pinpoint if any of the patient characteristics might help to determine the optimal depression treatment for each patient.

The study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. For detailed study design and primary analysis see Keller et al. [27]. We considered $p = 61$ baseline covariates for our observation matrix X; these variables are listed in Table 4.3. The outcome, R, was the 24-item Hamilton Rating Scale for Depression score [22], observed post treatment. For simplicity, we only allowed the action to vary between two treatments at a time. Since the primary analysis of the data showed the combination treatment to be superior to either individual treatment alone, we ran the variable selection techniques twice: the first time with the action varying between the combination treatment and Nefazodone alone, and the second time with the action varying between the combination treatment and CBASP alone.

The results of our first analysis comparing the combination treatment to Nefazodone alone are shown in Table 4.3 and Figure 4.3. The adjusted selection percentages for each variable are listed in Table 4.3 along with 80% and 90% thresholds at the bottom (i.e. alpha equal to 0.2 and 0.1). Figure 4.3 shows plots of these

adjusted selection percentages and thresholds. The x-axis in each plot corresponds to the variable numbers listed in Table 4.3. The horizontal dashed lines are 80% thresholds and the horizontal solid lines are 90% thresholds.

Only the adjusted selection percentages from the bootstrap resampling of CV Lasso are plotted in the first plot. Absent a working threshold, it is not very clear which variables should be selected for further analysis. The next three plots are for BIC Lasso, method U and method S. All three of these methods had one variable with an adjusted selection percentage exceeding the 80% threshold. For BIC Lasso, this variable was variable 34, *Obsessive Compulsive Disorder*, whereas, for both of the new methods the variable was variable 38, past history of *Alcohol Dependence*. Further analysis of the two variables confirmed that the interaction with *Obsessive Compulsive Disorder* and the action was non-qualitative in the data, whereas the interaction between past *Alcohol Dependence* and the action had good potential for being qualitative. More study should be done to determine the usefulness of past history of *Alcohol Dependence* for selecting treatments. Also, in this study around 20% of subjects in each group left the study early. Here R is the last observed Hamilton Rating, which is a worst case scenario under the assumption that depressed subjects who drop out of the study do not improve. Although we included all available good predictors of R in the model of $E[R|X, A]$, it may be the case that unobserved determinants of dropout provide an alternate explanation for the apparent qualitative interaction with past *Alcohol Dependence*.

The results of our second analysis comparing the combination treatment to CBASP alone are shown in Figure 4.4. The figure shows plots of the adjusted selection percentages for each method along with 80% thresholds. The x-axis in each plot corresponds to the variable numbers listed in Table 4.3. The horizontal dashed lines are

Table 4.3: Results from variable selection techniques on the Nefazodone CBASP trial data comparing the combination treatment against Nefazodone alone

	Adjusted Selection Percentages			
Variable	**CV Lasso**	**BIC Lasso**	**Method U**	**Method S**
1 Gender	43.1	8.3	4.0	3.4
2 Racial Category	6.2	1.5	0.4	0.9
3-4 Marital Status	12.3,12.4	0.3,1.6	0.1,0.3	0,0.6
5 Body Mass Index	2.2	1.2	1.0	0.8
6 Age in Years at Screening	20.8	2.6	1.8	1.1
7 Family/Friend Support System	2.3	0.7	0.3	0.1
8 Treated Current Depression	5.4	0.6	0.1	0.2
9 Psychotherapy Current Depression	28.5	2.9	1.3	1.2
10 Medication Current Depression	31.8	4.1	0.4	0.8
11 Treated Past Depression	35.1	18.7	3.5	3.8
12 Psychotherapy Past Depression	53.7	33.1	5.2	6.0
13 Medication Past Depression	21.7	13.5	1.8	2.0
14 Age of MDD Onset	12.8	3.5	2.2	2.0
15-17 Depressive Episodes Count	23.6,37.7,14.6	4.3,2.4,2.7	0.4,1.6,1.5	1.0,0.7,0.9
18 Length Current episode	38.0	3.6	1.5	0.4
19-20 MDD Current Episode Type	29.3,35.5	2.5,3.7	5.4,5.5	5.2,6.3
21-22 MDD Current Severity	20.5,9.3	1.1,0.6	2.9,1.8	3.6,1.3
23 Dysthymia Onset	27.6	1.3	0.1	0.1
24 Length Current Dysthymia	15.7	1.9	0.2	0.1
25-26 Generalized Anxiety	34.8,13.8	10.6,0.9	2.8,1.5	4.2,2.7
27 Anxiety Disorder NOS	49.3	18.9	1.1	0.5
28-29 Panic Disorder	35.2,38.8	5.4,18.3	1.3,3.7	0.4,3.5
30-31 Social Phobia	5.1,41.3	1.1,7.0	1.4,2.3	0.4,1.8
32-33 Specific Phobia	6.0,36.1	0.1,10.6	0.8,7.3	0.2,11.6
34 Obsessive Compulsive	59.8	**47.3**	12.6	12.6
35 Body Dysmorphic Current	23.9	2.1	2.2	3.6
36 Anorexia or Bulimia Nervosa	12.9	0.0	0.7	0.3
37-38 Alcohol Abuse/Dependence	48.3,62.0	19.0,35.4	25.4,**45.7**	24.1,**44.9**
39 Drug Abuse	1.9	3.1	1.1	1.1
40-41 Post Traumatic Stress	2.7,16.4	2.1,2.9	0.7,0	0.1,0.1
42 Other Psychological Problems	21.9	5.7	2.6	2.7
43 Global Assessment of Function	9.5	2.6	2.0	1.2
44-45 Main Study Diagnosis	3.8,36.9	0.3,6.0	0.7,2.3	0.2,2.8
46 Severity of Illness	9.8	0.7	0.1	0.2
47 Total HAMA Score	22.9	8.0	9.7	5.4
48 HAMA Sleep Disturbance	10.8	0.7	0.1	0.2
49 HAMA Psychic Anxiety Score	6.4	0.4	1.0	0.4
50 HAMA Somatic Anxiety Score	57.1	26.9	30.3	23.0
51 Total HAMD-24 Score	3.0	0.6	0.3	0.0
52 Total HAMD-17 Score	20.3	0.6	0	0
53 HAMD Cognitive Disturbance	2.0	0	0.7	0.1
54 HAMD Retardation Score	2.7	0.2	0.1	0.2
55 HAMD Anxiety/Somatic	2.1	0.3	0.3	0.1
56 IDSSR Total Score	8.8	0.3	0.2	0
57 IDSSR Anxious Depression Type	6.7	0.3	0	0
58 IDSSR General/Mood Cognition	23.4	6.3	4.4	2.7
59 IDSSR Anxiety/Arousal Score	4.3	0.1	0.1	0.1
60-61 IDSSR Sleep Scores	22.8,9.1	2.5,0.7	1.4,0.5	0.9,0.2
Thresholds: 80%, 90%	NA	**39.8, 50.3**	**42.4, 45.6**	**41.2, 46.5**

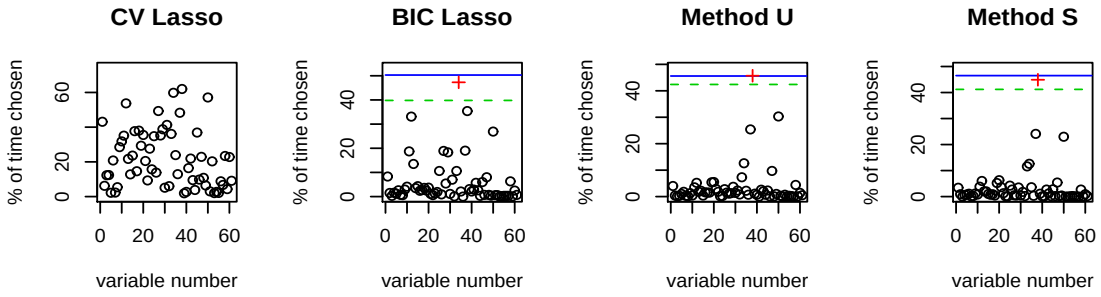


Figure 4.3: Plots of interaction variables selected from Nefazodone CBASP trial data comparing the combination treatment to Nefazodone alone. In each plot x-axis is the variable number given in Table 4.3, and y-axis is adjusted percent of time the variables were selected by the method. Dashed horizontal line is the 80% threshold and solid horizontal line is the 90% threshold. In the second plot the red + identifies the *Obsessive Compulsive Disorder* variable, whereas the red + in the third and fourth plots denotes *Alcohol Dependence*

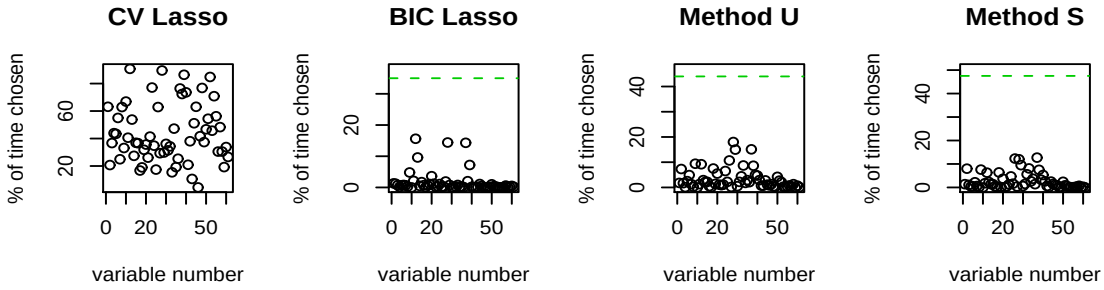


Figure 4.4: Plots of interaction variables selected from Nefazodone CBASP trial data comparing the combination treatment to CBASP alone. In each plot x-axis is the variable number given in Table 4.3, and y-axis is adjusted percent of time the variables were selected by the method. The dashed horizontal lines are 80% thresholds.

80% thresholds. As shown in the plot, no variables were selected by either of the new methods or BIC Lasso. For brevity we forgo listing the individual selection percentages. This analysis suggests there are no true qualitative interactions between the baseline covariates and the two treatment options. Many researchers believe this is the most likely scenario in medical decision making applications. We conclude that the combination treatment is better than CBASP alone for all patient subsets tested.

4.6 Discussion

In this chapter, we discussed when a variable is important in decision making and why variable selection techniques designed for prediction may not perform well in a decision making setting. We presented two new techniques explicitly designed to select variables for decision making. These techniques focus on interaction variables that are good candidates for playing a role in the actual decision rules.

It should be noted that Lasso treats the indicator variables used to model a categorical variable as separate variables. It is well known that this can lead to over selection of categorical variables with many categories. Consequently, the proposed method is subject to this problem. Therefore we recommend using Group Lasso [63, 65] or something similar in step 3 of the algorithm when applying it to a data set with many multi-category variables.

The entire algorithm including bootstrap sampling and thresholding takes approximately 30 hours to run in Matlab on a 3 Ghz Intel Xeon X5355 processor for a data set of $p = 60$ baseline covariates and $n = 400$ subjects. The algorithm would require far less computation time if a more theoretical justified threshold could be determined, rather than using a permutation based threshold. This is an area for future work.

More research is needed to determine the oracle consistency properties of this algorithm and its performance on problems where $p > n$. Adjusting these methods to deal with dropout is also an open issue. Our long term goal is to extend these methods to settings with multiple decision time points.

CHAPTER V

Variable Selection for Qualitative Interactions While Controlling the Family-wise Error Rate

5.1 Introduction

While the main goal of most clinical studies is to determine an overall optimal treatment, a critical question often asked is whether this overall optimal treatment is the best treatment to prescribe across different patient subsets. Many studies have pre-specified patient subsets they plan to test for differences in treatment effect. Others lack the expertise to know which patient characteristics play a critical role in the effectiveness of different treatments.

The topic of patient subset analysis has seen a good deal of attention throughout the last 30 years [9, 42, 50, 17, 64, 2, 48, 62, 30], a large amount of it seemingly controversial. The way clinical trials are designed and the nature of the topic makes the task difficult. Nevertheless, many clinicians feel these types of analysis are worth while and continue to seek out better ways to determine which treatments are best for individual patients.

In this chapter we focus on a specific type of subgroup analysis that indicates there should be a change in the choice of optimal treatment for certain subgroups. This occurs when a qualitative interaction exists between treatment and at least one patient characteristic. We propose a method for finding these qualitative interactions

in situations when prior intuition is lacking. We ensure the method also maintains small susceptibility to finding spurious results.

5.2 Qualitative Interactions

We consider subset analysis in the simplest setting where one must decide between two treatments. Let $X = (X_1, X_2, ..., X_p)$ be covariate observations about a subject and let A represent the treatment action. If the response to the treatment is labeled R, then the goal in most clinical studies is to find the treatment a^* for which

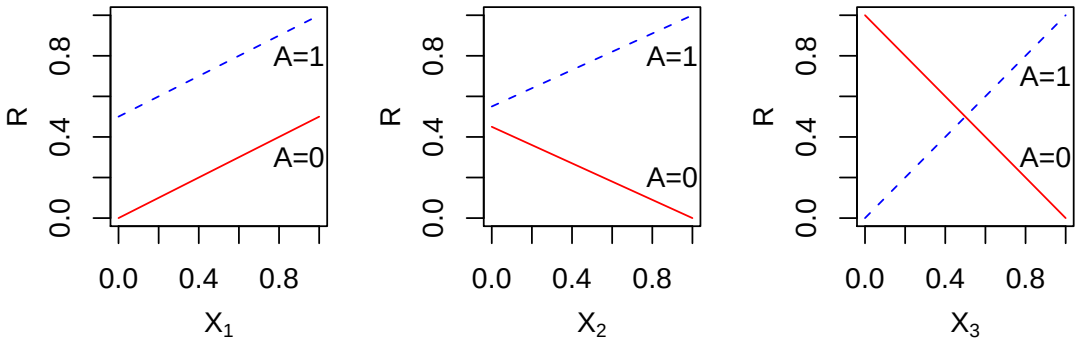
$$a^* = \arg\max_a E[R|A = a]. \tag{5.1}$$

Treatment effect varies across different subgroups of patients when the the treatment interacts with a covariate. Some types of interactions are more important than others when determining optimal treatments. Peto describes these types of interactions as qualitative interactions [42]. A variable X_j qualitatively interacts with the treatment, A, if there exists at least two distinct, non-empty sets, $S_1, S_2 \subset space(X_j)$ for which

$$\arg\max_a E[R|X_j = x_{j1}, A = a] \neq \arg\max_a E[R|X_j = x_{j2}, A = a],$$

for all $x_{j1} \in S_1$, and $x_{j2} \in S_2$. These variables are useful for prescribing treatment since they help decipher which treatment is optimal for different subsets of patients.

To illustrate this idea, see the plots in Figure 5.1. These plots depict different possible relationships between the conditional mean of R, A and a particular X_j, when averaging over all other X_i, $i \neq j$. Figure 5.1(a), shows a variable, X_1, which does not interact with the action. Figure 5.1(b) shows a variable, X_2, that interacts with the action, A, but does not qualitatively interact with the action. In both plots, the optimal action is $A = 1$. Figure 5.1(c), shows a variable, X_3, which qualitatively



(a) No interaction (b) Non-qualitative interaction (c) Qualitative Interaction

Figure 5.1: Plots demonstrating qualitative and non-qualitative interactions

interacts with the action. This type of interaction is more important since it impacts the best choice of treatment.

Much of the statistical literature suggests that the search for qualitative interactions should be limited to only pre-specified covariates and qualitative interactions that are found should be initially mistrusted [42, 30, 64, 48]. This point of view is understandable given the large number clinical trials claiming discovery of new qualitative interactions which are later refuted. However, it severely limits the ability of clinicians to make new scientific discoveries that may be critical to the practice of medicine. A better approach is to develop methods which increase the power to find qualitative interactions yet minimize the finding of spurious results.

There are currently a few qualitative interaction tests that can be used to test a small number of pre-specified interactions [50, 17, 39, 51, 38, 29, 61, 62, 45]. When controlling the error rate for multiple testing, these tests are quite conservative if the set of interactions being tested is large [17, 43, 62]. Also, many tests were designed to test for only qualitative interactions between categorical variables and the treatment action. In the next section we present a new method for finding qualitative interactions that demonstrates better power than current methods yet

also limits the false discovery rate.

5.3 AGV Lasso

The search for qualitative interactions involves the comparison of multiple strategies for choosing treatment. These strategies for choosing treatments are often referred to as policies or treatment regimes. A policy, π, is just a stochastic or deterministic decision rule mapping the space of observations, X, to the space of the treatment action, A. In other words, π defines the probability for choosing treatment action $A = a$ given the observations $X = x$.

We compare policies via the expected mean response, called the Value of a policy [53]. Let the distribution of X be a fixed distribution f, and let the distribution of R given (X, A) be a fixed distribution g. Then when actions are chosen according to a policy π, the trajectory (X, A, R) has distribution

$$f(x)\pi(a|x)g(r|x,a),$$

If $E_\pi[]$ denotes the expectation over the above distribution, then the Value of π is

$$V_\pi = E_\pi\left[R\right]$$

The optimal policy, π^*, is defined as

$$\pi^* = \arg\max_\pi V_\pi = \arg\max_\pi E_\pi\left[R\right],$$

or equivalently

$$\pi^*(x) = \arg\max_a E\left[R|\, X = x, A = a\right].$$

Our variable selection algorithm focuses on the change in Value of the estimated optimal policy when a variable is added to the model:

$$\max_a \hat{E}\left[R|X_j = x_j, A = a\right] - \hat{E}\left[R|A = a^*\right] \tag{5.2}$$

where $a^* = \arg\max_a E[R|A = a]$. Parmigiani refers to this quantity as the value of information [40].

The following is an overview of the algorithm.

Variable Selection Algorithm

1. **Rank the variables:** Rank the variables in $(X, A*X)$ using a Lasso. Define the variable rank to be the order in which the Lasso coefficients become non-zero.

2. **Create nested subsets of variables:** Create $2p$ nested subsets of the variables based on the rank order of the $2p$ variables in the previous step. Be sure to include the main effect variables of all interactions in the subset to satisfy the hierarchical ordering principle.

3. **Select between subsets using Adjusted Gain in Value Criterion:**

   (a) For each subset $k = 1, ..., 2p$, estimate the maximal Value, e.g.

      i. Use the subset to estimate $\hat{E}$

      ii. Estimate the optimal policy, $\hat{\pi}_k^*(x) = \arg\max_a \hat{E}[R|X = x, A = a]$

      iii. Estimate the Value of $\hat{\pi}_k^*$ by:

$$\hat{V}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{E}[R|X = x_i, A = \hat{\pi}_k^*(x_i)]$$

   (b) Select the subset, k^*, that has the highest Adjusted Gain in Value (AGV) criterion:

$$AGV_k = \frac{\hat{V}_k - \hat{V}_0}{\hat{V}_m - \hat{V}_0}\left(\frac{m}{k}\right)$$

   where $m = \arg\max_k \hat{V}_k$ and $\hat{V}_0$ is the estimated Value of the policy $\hat{\pi}_0^* = \arg\max_a \hat{E}[R|A = a]$.

In the first two steps we seek a quick way to navigate through the space of all possible combinations of the variables $(X, A*X)$. First we use Lasso [54] to rank the

variables. Lasso is a penalized regression procedure which returns a sparse, piecewise linear coefficient vector. It utilizes the L_1-norm of the coefficient vector, $|\beta|_1$, as its penalty function. The L_1-norm causes some of the coefficients to be set exactly to zero. We fit the Lasso on $(X, A*X, A)$, but leave the coefficient of A unconstrained by the L_1 penalty function. The rankings for the variables in $(X, A*X)$ are determined based on the order the variables enter the Lasso model. These rankings are then used to create nested subsets of the variables.

We rank all of the variables in the $(X, A*X)$, including the main effects, X, because they may be strongly predictive of the response variable, R, and will help reduce variability in the estimations. Also, when testing for the interaction between X_j and A, researchers often prefer to maintain a hierarchical ordering [59] and thus the main effect of the variable X_j are included. This helps to avoid finding spurious interactions that may appear because the main effect is important but is not included in the estimation.

However, Lasso favors variables that are predictive, so we offset this by using the Adjusted Gain in Value (AGV) criterion to select the optimal subset. The AGV criterion trades off between the complexity and the observed Value of each of the models. The criterion selects the subset of variables with the maximum proportionate increase in Value per variable. It is similar in idea to the adjusted R^2 value. The model with $m = \arg\max_k \hat{V}_k$ variables is akin to a saturated model, because the addition of more variables does not improve the Value of the model. Thus the denominator is the observed maximum gain in value, among the different variable subsets, divided by m, an estimate of the degrees of freedom used to achieve that gain in Value. The numerator then measures the gain in Value of the intermediate model, the model with k variables, divided by k, the estimated degrees of freedom

needed to achieve that gain in Value.

5.4 Controlling the familywise error rate

The familywise error rate (FWER) is the probability of making at least one false discovery among all hypothesis when performing multiple testing procedures [58, 49]. In this case the FWER is then the probability of selecting at least one spurious qualitative interaction among all interaction variables being considered in our variable selection procedure.

It may be acceptable in some instances to disregard the FWER when testing for a qualitative interaction between treatment and a small number of pre-specified variables. Controlling just the per test error rate may be sufficient for the desired analysis. When performing a large number of hypothesis tests, however, it becomes a necessity to employ some method which adjusts for the multiplicity of testing to control the FWER. This is the case with variable selection, and in particular, variable selection for qualitative interactions. Naturally these multiplicity correction methods decrease the power to find qualitative interactions. The failure to incorporate these method in the variable selection process, however, may result in wasted resources and weakened credibility. We illustrate this issue in the next section.

We suggest a combination of bootstrap sampling and permutation thresholding to help control the FWER when using the algorithm proposed in Section 5.3. First we use bootstrap sampling [14] of the original data to give a measure of reliability on the variables selected. The bootstrap samples allow us to determine the percentage of time each interaction variable is selected by the method. These selection percentages, with a slight adjustment, can be thought of as pseudo test statistics for each interaction variable. We compute the adjusted selection percentages for each

variable as follows.

1. Take 1000 bootstrap samples of the original data
2. Run variable selection algorithm and record the interaction variables that are selected along with the sign of the interaction coefficient for each bootstrap sample
3. Calculate the adjusted selection percentage across the 1000 bootstrap samples for each interaction variable: the absolute value of the number of times the interaction is selected with a positive coefficient minus the number of times an interaction is selected with a negative coefficient

This adjustment used in step 3 helps eliminate variables that, across the bootstrap samples, do not consistently interact in one direction with the action.

Second, we construct a permutation threshold to control for the number of false discoveries and determine which interaction variables to include in the final model. The threshold estimates the selection percentages we would expect to see if the data contained no interactions. To compute the permutation threshold:

1. Permute the X values of the $X * A$ interactions in the $(X, A, X * A)$ model matrix 100 times
2. On each permuted data set
   (a) Take 1000 bootstrap samples of the permuted data
   (b) Run variable selection algorithm and record the interaction variables that are selected along with the sign of the interaction coefficient for each bootstrap sample

(c) Calculate the adjusted selection percentage across the 1000 bootstrap samples for each interaction variable: the absolute value of the number of times the interaction is selected with a positive coefficient minus the number of times an interaction is selected with a negative coefficient

(d) Record the maximum selection percentage observed across the p interaction variables

3. Define the permutation threshold to be the $(1-\alpha)th$ percentile over the 100 maximum selection percentages for each permuted data set

We chose all interaction variables whose adjusted selection percentage from the original data is greater than the permutation threshold.

Permutation-based multiplicity correction procedures are discussed in detail by Westfall and Young [58]. They have seen widespread use and success in many scientific applications such as microarray analysis and medicine and even variable selection for prediction [33, 13, 52, 55].

In the next section we show simulation results testing the proposed variable selection algorithm with permutation threshold. We reference this method as AGV Lasso.

5.5 Size and Power Comparisons

We ran AGV Lasso on realistically designed simulation data to test its performance and compared the results to two different methods suggested for formally testing for qualitative interactions.

In order to generate realistic simulation data, we randomly selected rows, with replacement from X, the observation matrix from the Nefazodone CBASP trial data. We generated new actions, A, and new responses, R, that covered a wide variety of

models. We report results for the following generative models:

1. Main effects of X only, no treatment effect and no interactions with treatment
2. Main effects of X, moderate treatment effect and no interactions with treatment
3. Main effects of X, moderate treatment effect, multiple small non-qualitative interactions with treatment, no qualitative interaction with treatment
4. Main effects of X, moderate treatment effect, multiple moderate non-qualitative interactions with treatment, no qualitative interaction with treatment
5. Main effects of X, small treatment effect, small qualitative interaction with a binary variable, no non-qualitative interactions
6. Main effects of X, small treatment effect, small qualitative interaction with a continuous variable, no non-qualitative interactions
7. Main effects of X, small treatment effect, multiple small non-qualitative interactions with treatment, small to moderate qualitative interaction with a binary variable and treatment
8. Main effects of X, small treatment effect, multiple small to moderate non-qualitative interactions with treatment, small qualitative interaction with a continuous variable and treatment

For each generative model, we used main effect coefficients for the variables X, estimated in an analysis of the real data set. In generative models 3-7 we randomly selected variables from the Nefazodone CBASP data for each treatment covariate interaction and used these same variables for each repetition. The treatment, qualitative interaction and non-qualitative interaction coefficients were set using a variant

of Cohen's D effect size measure [10] shown below:

$$D = \frac{\beta\sqrt{Var(R)}}{\sqrt{Var(X_j)}} \tag{5.3}$$

We altered this formula by replacing the marginal variance, $Var(R)$, with the conditional variance of the response $Var(R|X, A)$. However, we maintained the definitions of 'small' and 'moderate' effect sizes suggested by [10] as $D = 0.2$ and $D = 0.5$ respectively. Thus the effects are slightly smaller than the traditional definition.

We compared AGV Lasso to the likelihood ratio test (LRT) proposed by Gail and Simon [17]. The LRT is designed to test for a qualitative interaction between a binary treatment and a single categorical variable or a combination of categorical variables. Let δ_i, $i = 1, ..., I$ be the true treatment effects for each of the I categories of subjects and let D_i, $i = 1, ..., I$ be independent normal estimates of those effects with variances σ_i^2. Define

$$Q^+ = \sum_{i=1}^{I} \frac{D_i^2}{\sigma_i^2} I(D_i > 0) \tag{5.4}$$

and

$$Q^- = \sum_{i=1}^{I} \frac{D_i^2}{\sigma_i^2} I(D_i < 0) \tag{5.5}$$

The LRT for testing the null hypothesis that $\delta_i \geq 0$ for all i or $\delta_i \leq 0$ for all i is then

$$T_Q = min(Q^+, Q^-) > c \tag{5.6}$$

where the constant c is chosen to ensure a significance level α. Gail and Simon [17] give several values of c for different I and α. The σ_i^2 in Equations 5.4 and 5.5 above can be replaced by a consistent estimate in large samples. Also, continuous variables must be dichotomized when using this test.

We also compared AGV Lasso to the qualitative interaction test proposed by Shuster and Van Eys [50]. This test is based on joint confidence intervals and can be

used to test for a qualitative interaction between a binary treatment and any type of covariate(s). Assume our response R is a linear function of the treatment and the covariates. For example it might be

$$R = \beta_0 + X_j\beta_1 + A\beta_2 + AX_j\beta_3 + \epsilon, \tag{5.7}$$

where ϵ is an error term. The treatment difference for subjects with $X_j = x_j$ would be $D(x_j) = \beta_2 + x_i\beta_3$. The parameter $-\beta_2/\beta_3$, is the value of X_j for which the treatments are equal. A asymptotic $(1-\alpha)\%$ confidence interval for $-\beta_2/\beta_3$ contains all values, x_j for which

$$(\hat{\beta}_2 + x_j\hat{\beta}_3)^2 < Z^2_{\alpha/2}(V_{22} + 2x_jV_{23} + x_j^2V_{33}) \tag{5.8}$$

where Z_α is the upper (100α) percent point of the standard normal curve and

$$V = \begin{bmatrix} V_{11} & V_{12} & V_{13} \\ V_{12} & V_{22} & V_{23} \\ V_{13} & V_{23} & V_{33} \end{bmatrix} \tag{5.9}$$

is the asymptotic covariance matrix of $\hat{\beta}$. All values falling in this confidence interval are values of X_j for which no significant treatment difference exists. The null hypothesis of no qualitative interaction is then rejected if the confidence interval for $-\beta_2/\beta_3$ is strictly contained in the range of X_j within the data. In other words the null hypothesis is rejected if there exists at least one x_{ij} in the range of X_j within the data for which there is a significant positive treatment effect and at least one x_{kj} in the range of X_j within the data for which there is a significant negative treatment effect. We can express this formally as

$$T_V = min(V^+, V^-) > Z^2_{\alpha/2} \tag{5.10}$$

where

$$V^+ = \max_{i=1,..,n} \frac{(\hat{\beta}_2 + x_{ij}\hat{\beta}_3)^2 I(x_{ij} > -\hat{\beta}_2/\hat{\beta}_3)}{(V_{22} + 2x_{ij}V_{23} + x_{ij}^2 V_{33})} \tag{5.11}$$

and

$$V^- = \max_{i=1,..,n} \frac{(\hat{\beta}_2 + x_{ij}\hat{\beta}_3)^2 I(x_{ij} < -\hat{\beta}_2/\hat{\beta}_3)}{(V_{22} + 2x_{ij}V_{23} + x_{ij}^2 V_{33})} \tag{5.12}$$

The test can also be modified to deal with multiple covariates (see [50]).

For each generative model, we ran AGV Lasso and the two qualitative interaction tests with and without corrections for multiplicity. We tried two multiplicity corrections for each qualitative interaction test. The first multiplicity correction method we tried was a Bonferroni correction due to its easy application with non-standard tests such as the LRT ([49]). This correction method tends to be conservative, however, so we also tried a permutation threshold similar to what we used in the new method. The permutation threshold was calculated in the same way except we replaced the selection percentages with the individual T-statistics (Equations 5.6 and 5.10) for each variable. We then selected all interaction variables whose T-statistic from the original data was greater than the permutation threshold.

We ran the analysis 200 times. We recorded the percentage of time each method selected one or more spurious interactions and the qualitative interaction (if one existed) to estimate the size and power of each method. The results are listed in Tables 5.1 and 5.2. The percentage of time one or more spurious qualitative interactions was selected by each method over the 200 repetitions is listed in Table 5.1. The percentage of time the true qualitative interaction was selected by each method over the 200 repetitions is listed in Table 5.2. Note that since generative models 1-4 have no qualitative interactions with treatment, power results are not applicable to these models.

Table 5.1: **Size Estimations:** The first two columns list the desired significance level and the method. AGVL stands for AGV Lasso, LRT stands for the Gail-Simon likelihood ratio test and SVE for the Shuster-Van Eys test, Bonferroni stands for a Bonferroni correction and permutation stands for the permutation based multiplicity correction. The last eight columns give give the percentage of time one or more spurious qualitative interactions was selected over the 200 repetitions for each generative model. Stared percentages fall outside the 95% confidence interval for the desired significance level

Sig.		**Generative Model**							
Level	**Method**	1	2	3	4	5	6	7	8
$\alpha = .05$	LRT uncorrected	13.0*	7.5	5.5	0.0	22.5*	17.0*	20.0*	14.5*
	SVE uncorrected	29.0*	9.5*	8.0	0.0	32.0*	35.5*	58.0*	31.0*
	LRT Bonferroni	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0
	SVE Bonferroni	1.0	0.0	0.0	0.0	0.0	0.5	4.0	0.0
	LRT permutation	6.0	4.5	3.0	0.0	8.5	7.5	11.0*	9.5*
	SVE permutation	3.5	6.0	6.5	0.0	6.0	8.0	21.5*	6.0
	AGVL	7.0	5.5	8.0	23.5*	7.5	7.0	3.5	6.0
$\alpha = .1$	LRT uncorrected	34.5*	15.0*	8.5	0.0	46.0*	38.5*	45.5*	31.0*
	SVE uncorrected	54.5*	29.5*	29.5*	1.0	65.5*	62.0*	74.0*	56.0*
	LRT Bonferroni	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0
	SVE Bonferroni	1.5	0.5	0.0	0.0	0.0	1.5	7.5	0.5
	LRT permutation	7.0	8.5	7.0	0.0	11.5	10.0	15.0*	11.5
	SVE permutation	6.5	7.5	9.0	0.0	8.5	11.5	26.5*	8.0
	AGVL	11.0	9.0	14.0	32.0*	10.5	11.5	5.5	10.5

Looking over Table 5.1 we see that without the multiplicity correction, the two test methods have large Type I error rates. The Bonferroni correction method is far more conservative than the permutation based multiplicity correction. AGV Lasso appears to maintain the desired FWER in all settings but one. Under generative model 4 AGV Lasso fails to maintain the desired significance level. Upon closer examination we discovered the failure was due to over selection of true non-qualitative interactions. This may be due to the fact that the permutation threshold targets all interactions as opposed to only targeting qualitative interactions.

Table 5.2 shows that the LRT is better suited to find qualitative interactions with a categorical covariate, as would be expected. Whereas, the Shuster-Van Eys test is much better at finding qualitative interactions with a continuous covariate. The new method seems to have good comparative power against the methods which control for the FWER.

Table 5.2: **Power Estimations:** The first two columns list the desired significance level and the method. AGVL stands for AGV Lasso, LRT stands for the Gail-Simon likelihood ratio test and SVE for the Shuster-Van Eys test, Bonferroni stands for a Bonferroni correction and permutation stands for the permutation based multiplicity correction. The last 4 columns give the percentage of time the true qualitative interaction was selected over the 200 repetitions for each generative model. Bolded percentages correlate with settings where the desired significance level was maintained.

Sig.		**Generative Model**			
Level	**Method**	5	6	7	8
$\alpha = .05$	LRT uncorrected	12.0	8.5	52.0	12.5
	SVE uncorrected	9.0	24.0	45.5	55.0
	LRT Bonferroni	**0.5**	**0.5**	**8.5**	**0.0**
	SVE Bonferroni	**0.0**	**1.0**	**4.5**	**8.5**
	LRT permutation	**6.5**	**5.0**	34.0	7.5
	SVE permutation	**0.5**	**3.0**	15.0	**24.0**
	AGVL	**14.0**	**13.0**	**44.0**	**20.5**
$\alpha = .1$	LRT uncorrected	21.5	13.0	59.0	20.5
	SVE uncorrected	18.0	33.0	58.0	66.5
	LRT Bonferroni	**0.5**	**0.5**	**8.5**	**0.0**
	SVE Bonferroni	**0.5**	**2.5**	**8.5**	**12.5**
	LRT permutation	**7.5**	**7.0**	41.0	**8.0**
	SVE permutation	**1.0**	**3.5**	19.0	**30.0**
	AGVL	**17.5**	**16.5**	**49.0**	**26.5**

Overall, we found that the new method performs better than the other two tests when controlling for the FWER. While, the competing methods each have there individual strengths, they seem to lack consistent performance to merit use as a generalized variable selection method for qualitative interactions.

5.6 Example

We applied AGV Lasso along with the LRT test and the Shuster-Van Eys test to the Nefazodone CBASP trial [27] data introduced in Chapter IV. This trial was conducted to compare the efficacy of three alternate treatments for patients with chronic depression. The study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. We considered the same $p = 61$ baseline covariates for our observation matrix X listed in

Table 4.3. The outcome, R, was the 24-item Hamilton Rating Scale for Depression score [22], observed post treatment. In this analysis we only look at a subset of the study consisting of the $n = 440$ patients who were randomized to either the combination treatment or Nefazodone alone.

Using a 90% permutation threshold, AGV Lasso selected two variables. Both variables had the same selection percentage of 21.9%, which was slightly higher than the 90% threshold of 21.1%. These variables were *Obsessive Compulsive Disorder* and past history of *Alcohol Dependence*. No variables were selected by the qualitative interaction tests using either multiplicity correction at $\alpha = 0.1$.

5.7 Discussion

Although multiple tests exist for evaluating qualitative interactions, they are designed to be used on a small number of covariates, often of a particular form. We have proposed a new technique that can be used to find qualitative interactions among a large number of covariates. We have included measures to ensure the FWE error rate is controlled for, an important characteristic for methods used in post-hoc analysis. The methods proposed here can be used with multiple different types of covariates without predetermining the best division into subsets.

In the future we hope to modify the way we permute the data in the permutation threshold so that it targets just the qualitative interactions instead of all interactions. We believe this would eliminate the over selection of non-qualitative interactions in data similar to generative model 4. We also think it would be useful to try replacing the Lasso in the algorithm with other types of penalized regression models to allow for different types of response variables such as binary or survival. Our ultimate goal, however, it to develop a variable selection method for sequential decision making

applications like SMART trials [36].

CHAPTER VI

Suggestions for Further Research

Chapters II and III gave a background on decision making applications and current techniques used for variable selection. Chapters IV and V presented ideas for variable selection in single time point decision applications. This chapter will discuss ideas for improving the methods presented in Chapters IV and V. It will outline key issues for expanding these methods to sequential decision making problems and give ideas for addressing these issues in future research.

6.1 Value Based Rankings

All of the methods presented in this thesis rely on variable selection techniques designed for prediction. Predictive variables are important for estimating models in a decision making process. Predictive variables also help reduce the variability of estimates used to find qualitative interactions. We may see improvements in our method, however, by minimizing or eliminating its dependence on predictive variable selection methods.

In particular, the variable selection algorithm presented and tested in Chapter V used a predictive variable selection method, Lasso, to rank the variables in the first step. We may see an increase in power to select the qualitative interaction if the Lasso ranking is replaced with a ranking procedure which ranks variables according

to their ability to affect the Value function.

We have attempted this in our research, but have not yet been successful at increasing the power. We outline our attempt below.

6.1.1 Classification Formulation for One Time Point Decision Making

Computer science researchers proposed a group of techniques for finding optimal policies by formulating decision making problems as classification problems [31, 32]. Formulating the decision making problem as a classification problem allows the use of a large number of sophisticated techniques developed to solve classification problems. We briefly describe this formulation.

In decision making we wish to find a non-stochastic decision rule $d : \mathcal{X} \rightarrow \mathcal{A}$ such that the expected reward, R, when using that decision rule is maximized over the population of possible $\mathbf{X}$, A and R values. In other words, if the distribution of $\mathbf{X}$ is $g(\mathbf{x})$, the distribution of A given $\mathbf{X}$ is $p(a|\mathbf{x})$ and the distribution of R given $\mathbf{X}$ and A is $h(r|\mathbf{x}, a)$, we want to find d such that

$$\max_{d(\mathbf{X})} E\left[\frac{1(d(\mathbf{X})=A)}{p(A|\mathbf{X})}R\right] = \max_{d(\mathbf{x})} \int r1\{d(\mathbf{x}) = a\}g(\mathbf{x})h(r|\mathbf{x}, a). \tag{6.1}$$

This optimization problem is equivalent to a weighted classification problem where we classify patients by their actions, A, based on $\mathbf{X}$, with the importance of each sample observation i is given by $\frac{r_i}{p(a_i|\mathbf{x}_i)}$.

In this formulation, we are only modeling the interactions between the action and the covariate observations X, so we lose some information about the decision making process. However, the methods have been used successfully on a few problems. We decided to test this formulation out in our variable selection problem since there are several good techniques for variable selection in classification. We tested a weighted L_1-norm penalized support vector machine using this formulation.

6.1.2 Interaction Ranking by L_1-Norm Penalized Support Vector Machine

In a classification setting where the outcome variable is binary $Y \in \{-1, 1\}$ and we use a model $f(\mathbf{x}) = \beta_0 + \mathbf{x}\beta$ to estimate y by $\hat{y} = sign(f(\mathbf{x}))$, the classical linear support vector machine (SVM) classifier solves the following regularization problem:

$$\min_{\beta_0,\beta} \tfrac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{n} H_1(y_i(\beta_o + \mathbf{x}_i\beta)) \tag{6.2}$$

where C is a tuning parameter and $H_1(v) = (1-v)_+$ is the hinge loss with $(v)_+ = v$ if $v > 0$ and 0 otherwise. The SVM tries to classify the data by finding the $(p-1)$ dimensional hyperplane which maximizes the margin between the two classes.

Observation weights can be easily incorporated into the optimization by replacing C in the above equation with Cw_i. If we replace the L_2-norm in Equation 6.2 with the $L1$-norm we get the L_1-norm penalized linear SVM [66]. The weighted L_1-norm penalized linear SVM is then

$$\min_{\beta_0,\beta} ||\beta||_1 + C\sum_{i=1}^{n} w_i H_1(y_i(\beta_o + \mathbf{x}_i\beta)) \tag{6.3}$$

This optimization problem can be solved using linear programming.

Using the classification formulation for decision making, with binary A, we get the following for equation 6.3:

$$\min_{\beta_0,\beta} ||\beta||_1 + C\sum_{i=1}^{n} \frac{r_i}{p(a_i|\mathbf{x}_i)} H_1(a_i(\beta_o + \mathbf{x}_i\beta)) \tag{6.4}$$

and $\hat{a}(\mathbf{x}_i) = sign(f(\mathbf{x}_i)$.

We used the weighted L_1-norm linear SVM to rank the variables based on their importance in classifying subjects by their optimal action. The weighted classifiers attempt to model the interaction relationships between X,A and R. When a variable appears important to the classifier, it can be interpreted that the interaction between X and A is important to the maximization of R.

Like the Lasso used in the algorithms in Chapters IV and V, the L_1-norm penalization of the linear SVM causes the solution vectors for β to be sparse when C is small. This allowed us to replace the Lasso in AGV Lasso in a similar fashion with the weighted L_1-norm SVM.

We tried the following algorithm:

1. Rank the interaction variables in X using the weighted L_1-norm linear SVM
2. Create p nested subsets of the interaction variables based on the order of entry of the p variables in the weighted $L1$-norm penalized classifier used in the previous step; include main effect variables as needed to satisfy the hierarchical ordering principle
3. Select the best subset using AGV criterion

We tested the algorithm on simulated data and unfortunately found the classifiers had a difficult time highly ranking the true qualitative interaction. We suspect this occurs because we are only modeling the interactions in the classification formulation and this can lead to high variability in the estimation.

6.1.3 Alternate Ideas

Another idea for value based rankings is to use something similar to the S-score presented in Section 4.3. Predictive variables should be incorporated in some way into the estimation portion of the algorithm. Care is needed when using the S-score in conjunction with the AGV criterion to avoid over fitting, given the similarities between the two.

6.2 Improving the Permutation Thresholds

In Chapter IV we introduced permutation thresholds for controlling the family-wise error rate. We tested the permutation thresholds in Chapter V and found that they effectively control the family-wise error rate in all settings except for models with moderate to large non-qualitative interactions but no qualitative interactions.

We find the thresholds by permuting the X values of the $X * A$ interactions in the $(X, A, X * A)$ data matrix. This permutation of the data removes the effects of all interactions, not just the qualitative interactions. We might see better size results in settings with only large non-qualitative interactions if we adjust the way we permute the data so that it targets qualitative interactions more than all other types of interactions. One possible approach is to only permute variables with moderate or high selection percentages across the bootstrap samples of the original data.

We also may observe an increase in power if we calculate a separate threshold for each individual interaction. Instead of permuting all the X values of the $X * A$ interactions at once, we would permute one variable in X at a time and re-run the analyses. This would require a far greater computational cost, especially when testing for size and power.

6.3 Performance With Non-binary Actions

Most of the methods presented in this thesis are easily generalized for use on problems with non-binary action spaces. Some minor adjustments may need to be made. The methods have not been tested in this setting and it would be useful to know how well they perform as the number of possible actions increases.

6.4 Dealing With Multiple Decision Time Points

Since the setting most likely to benefit from variable selection is the sequential decision making problem, this section will highlight some of the issues that arise in variable selection when there are multiple time points. This section also explains how the current single time point methods may be adjusted to deal with these issues.

6.4.1 Adjusting Variables that are Outcomes to Prior Actions

The variables observed after actions are taken, $X_2, ..., X_T$ can be considered outcomes to prior actions and may be affected by those prior actions. Thus, policies that differ from the policy used to collect the data, may change how $X_2, ..., X_T$ are distributed.

The distribution of a variable is important in determining the usefulness of the variable and the methods we have proposed. Recall Figure 4.2, plots 4.2(a) and 4.2(b) showed variables X_4 and X_5 which had similar relationships to the response variable and the action but their distributions were centered differently. Based on this information, X_4 appears more beneficial than X_5. So if the prior actions affect the distribution of $X_2, ..., X_T$, the variable selection methods may omit variables that are only useful when optimal prior actions are taken or include variables that are no longer useful when optimal prior actions are taken.

For example, suppose the Nefazodone CBASP trial introduced in Chapter IV was expanded to include a second treatment for patients who did not do well on their initial treatment. Before the second treatment is assigned, clinicians would collect data about the patient's condition after their initial treatment. One variable they might collect is a general measure of treatment burden. If a certain subset of patients found CBASP to be rather burdensome, the distribution of this variable would be

different for patients assigned the combination treatment than for patients assigned Nefazodone alone. If the optimal policy always chose the combination treatment for the initial treatment of future patients, this variable may need adjusted in order to determine its usefulness for prescribing the next treatment.

Thus, it is important to consider the distribution of $X_2, ..., X_T$ in sequential decision making problems and when necessary develop a way to adjust the distribution. A simple solution is to do a separate variable selection for each prior action. This approach is less appealing, however, with small sample sizes. A better approach may be to adjust the location and variance of the variables $X_2, ..., X_T$ for patients who were given suboptimal actions. Another way may be to up-weight subjects who received optimal actions and down-weight subjects who received suboptimal actions.

6.4.2 Adjusting the Outcome Variable

As discussed in Chapter II, at each decision time point we receive a response, R_t. Since actions may appear optimal under the current response despite being suboptimal under the sum of the responses, we must also consider future responses when modeling the current observation variables and actions. In the Q-learning algorithm (Section 2.2), the outcome used for modeling decision time points $t = 1, ..., T-1$ is the present response plus an estimate of the expected sum of future responses when following the optimal policy thereafter. This adjusted response helps determine which action is best over the sum of the responses. This adjusted response may more accurately represent the relationships in the decision making process concerning the optimal action, however, it may greatly increase the variability in the estimation. Research is needed to determine whether $R_t + R_{t+1} + .. + R_T$ or the adjusted response would work better for variable selection.

6.4.3 Order of Variable Selection

In Section 6.4.1 we discussed the possible need to adjust the distribution of predictors observed in time points $t = 2, ..., T$ based upon actions taken in the prior time points. This suggests doing variable selection moving forward through time. However, in Section 6.4.2, the outcome variable to use for variable selection in time points $t = 1, ..., T-1$, may need adjustment by the estimated sum of optimal future responses. This would suggest doing variable selection moving backward through time starting with time $t = T$ and ending with time $t = 1$. In essence, we want to do variable selection on variables obtained using optimal past actions and responses obtained using optimal future actions. Variable selection may need to be performed in an iterative fashion to satisfy both of these demands, iterating between adjusting the distribution of the observation variables and adjusting the estimates of the responses.

6.4.4 Troubles With Estimating the Value

The variable selection algorithms presented in Chapters IV and V depended upon estimates of the Value for derived optimal policies. We noted in Section 2.3.1, that as the policy being evaluated diverged from the policy used to collect the data, the variability in the estimated Value increased. This may limit our ability to use the estimated Value as a parameter selection criterion when T is large.

6.5 Conclusion

While there are many variable selection techniques designed for optimal prediction problems, this topic has had little attention in settings focused on optimal decision making. Techniques designed to find variables for decision making are rapidly becoming a necessity as the desire for evidence based decision making grows and the

ability to collect and store ever larger amounts of data increases.

In this thesis we highlighted some of the differences between prediction and decision making. We discussed the idea of qualitative interactions with the action and demonstrated why this characteristic delineates variables used for prediction versus variables used to make decisions.

We proposed multiple methods which capitalize on changes in the Value of optimal policies when interactions are added to the model. These new methods performed better in testing against competitive variable selection methods for prediction and commonly used tests for qualitative interactions.

In the future we hope to improve upon the methods and ideas expressed in this thesis. We also hope this work will spark other researchers to explore the topic and make advances in finding variables useful for sequential decision making.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] S. F. Assmann, S. J. Pocock, L. E.Enos, and L. E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, 355:1064–1069, 2000.

[3] R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Nashua, NH, 1995.

[5] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[6] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[7] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[8] P. J. Brown and M. Vannucci. Multivariate bayesian variable selection and predicting. *Journal of the Royal Statistical Society B*, 60:627–641, 1998.

[9] D. P. Byar and D. K. Corle. Selecting optimal treatment in clinical trials using covariate information. *Journal of Chronic Diseases*, 30:445–459, 1977.

[10] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Earlbaum Associates, Hillsdale, NJ, 1988.

[11] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 74–81, 2001.

[12] P. Dellaportas, J. J. Forster, and I. Ntzoufras. On bayesian model and variable selection using mcmc. Technical report, 2002.

[13] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistal Science*, 18:71–103, 2003.

[14] B. Efron and R. Tibshirani. *An introduction to the bootstrap*. Chapman and Hall, New York.

[15] D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.

[16] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96.

[17] M. Gail and R. Simon. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 41:361–372, 1985.

[18] E. I. George and R. E. McCulloch. Bayesian varable selection via gibbs sampling. *Journal of the American Statistical Association*, 88.

[19] E. I. George and R. E. McCulloch. Aproaches for bayesian varable selection. *Statistica Sinica*, 7:339–373, 1997.

[20] Beta-Blocker Pooling Project Reasearch Group. The beta-blocker pooling project (bbpp): Subgroup findings from randomized trials in post infarction patients. *The European Heart Journal*, 9:8–16, 1988.

[21] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[22] M. Hamilton. Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, 6:278–296, 1967.

[23] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, NY, 2001.

[24] S. D. Hollon and A. T. Beck. *Cognitive and cognitive behavioral therapies*, pages 447–492. Wiley, New York, NY, 2004.

[25] D. G. Horvitz and D. J. Thompson. Ageneralization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

[26] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

[27] M. B . Keller, J. P. McCullough, D. N. Klein, B. Arnow, D. L. Dunner, A. J. Gelenberg, J. C. Marekowitz, C. B. Nemeroff, J. M. Russell, M. E. Thase, M. H. Trivedi, and J. Zajecka. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for treatment of chronic depression. *New England Journal of Medicine*, 342:331–366, 2000.

[28] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[29] J. H. Krystal, J. A. Cramer, W. F. Krol, G. F. Kirket, and R. Rosenheck. Naltrexone in the treatment of alcohol dependence. *New England Journal of Medicine*, 345:1734–1739, 2001.

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

[31] M. Lagoudakis and R. Parr. Reinforcement learning as classification: Leveraging modern classifiers. *Proceedings of the 20th International Conference on Machine Learning*.

[32] J. Langford and B. Zadronzy. Relating reinforcement learning performance to classification performance. *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[33] F. Lindgren, B. Hansen, W. Karcher, M. Sjostrom, and L. Erik. Model validation by permutation tests: Applications to variable selection. *Journal of Chemometrics*, 10:521–532, 1996.

[34] M. Loth, M. Davy, and P. Preux. *Sparse Temporal Difference Learning using LASSO*. Hawaii, USA, 2006.

[35] K. P. Murphy. A survey of pomdp solution techniques. Technical report, 2000.

[36] S. A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, pages 1455–1481.

[37] S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society B*, 65:331–366, 2003.

[38] The ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in moderately hypercholesterolemic, hypertensive patients randomized to pravastatin vs usual care: The antihypertensive and lipid-lowering treatment to prevent heart attack trial(allhat-llt). *Journal of the American Medical Association.*

[39] G. Pan and D. A. Wolfe. *Statistics in Medicine*, 16:1645–1652, 1997.

[40] G. Parmigiani. *Modeling in Medical Decision Making: a Baysian Approach.* Wiley, West Sussex, England, 2002.

[41] L. Peskin. *Reinforcement Learning by Policy Searc.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.

[42] R. Peto. *Statistical Aspects of Cancer Trials*, pages 867–871. Chapman and Hall, London, UK, 1982.

[43] S. Piantadosi and M. H. Gail. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine*, 12:1239–1248, 1993.

[44] L. Portinale and L. Saitta. Feature selection. Technical Report IST-11993, 2002.

[45] C. F. Reynolds, M. A. Dew, B. G. Pollock, B. H. Mulsant, E. Frank, M. D. Miller, P. E. Houck, S. Mazumdar, M. A. Butters, J. A. Stack, M. A. Schlernitzauer, E. M. Whyte, A. Gildengers, J. Karp, E. Lenze, K. Szanto, S. Bensasi, and D. J. Kupfer. Maintenance treatment of major depression in old age. *New England Journal of Medicine*, 354:1130–1138, 2006.

[46] J. M. Robins. *Optimal Structural Nested Models for Optimal Sequential Decisions.* Springer, New York, NY, 2004.

[47] R. Y. Rubinstein. *Simulation and the Monte Carlo Method.* Wiley, New York, NY, 1981.

[48] S. J. Senn. Individual therapy: new dawn or false dawn.

[49] J. S. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

[50] J. Shuster and J. Van Eys. Interaction between prognostic factors and treatment. *Controlled Clinical Trials*, 4:2009–214, 1983.

[51] M. J. Silvapulle. Test against qualitative interaction: exact critical values and robust test. *Biometrics*, 57:1157–1165, 2001.

[52] R. M. Simon, E. L. Korn, L. M. McShane, M.D. Radmacher, G. W. Wright, and Y. Zhao. *Design and analysis of DNA microarray investigations.*

[53] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[54] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

[55] J. F. Troendle. Multiple comparisons between two groups on multiple bernoulli outcomes while accounting for covariates. *Statistics in Medicine*, 24:3581–3591, 2005.

[56] J. N. Tsitsiklis and B. Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94, 1996.

[57] C. J. Watkins. *Learning from Delayed Rewards.* PhD thesis, Cambridge University, Cambridge, 1989.

[58] P. H. Westfall and S. S. Young. *Resampling-based Multiple testing.* John Wiley and Sons, New York, NY, 1993.

[59] C. F. Wu and M. Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization.* Wiley, New York, NY, 2000.

[60] E. P. Xing and M. I. Jordan andR. M. Karp. Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 601–608, 2001.

[61] X. Yan. Test for qualitative interaction in equivalence trials when the number of centers is large. *Statistics in Medicine*, pages 711–722, 2004.

[62] X. Yan and X. Su. *Testing for Qualitative Interaction.* Informa Health Care, 2005.

[63] M. Yuan and Y. Lin.

[64] S. Yusuf, J. Wittes, J. Probstfield, and H. A. Tyrole. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *The Journal of the American Medial Association*, 266(1):93–98, 1991.

[65] P. Zhao, G. Roca, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. Technical Report 703, 2006.

[66] J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. 1-norm support vector machines. In *Neural Information Processing Systems*, page 16. MIT Press, 2003.

[67] H. Zou, T. Hastie, and R. Tibshirani. On the “degrees of freedom.