Subgroup Identification and Variable Selection from Randomized Clinical Trial Data

by

Jared C. Foster

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2013

Doctoral Committee:

Professor Bin Nan, Co-chair
Professor Jeremy M.G. Taylor, Co-chair
Assistant Professor Lu Wang
Professor Ji Zhu

For my wonderful family and beautiful fiancé.

I couldn't have done this without you.

ACKNOWLEDGMENTS

Thank you to Professors Jeremy Taylor, Bin Nan, Ji Zhu and Lu Wang, my dissertation committee, for all the very helpful questions and comments you have given me. Beyond this, I'd like to thank Jeremy and Bin, my co-advisors, for all the hours they spent meeting with me and all the useful advice they have given me. Bin, you introduced me to the incredibly fascinating field of variable selection, for which I am truly grateful. Jeremy, you introduced me to subgroup analysis and personalized medicine, which, combined with variable selection, has become my true passion. Additionally, despite my best efforts to become your most annoying student, you have been incredibly patient, thoughtfully responding to my many emails and discussing many ideas and issues with me whenever I randomly drop by your office. You guys are awesome, and I am truly blessed to have had the opportunity to learn from both of you.

Thank you also to my office mates, past and present, for putting up with me and my love of conversation. Yong-Seok Park, Rebecca Andridge and Connie Lee were incredibly helpful when I was preparing for the qualifying exam. Anna Conlon, John Rice, Jincheng Shen, Daniel Muenz and Oliver Lee, you have been very patient with me, and I have very much enjoyed our many fascinating conversations. Edward Kennedy, Phil Boonstra and Laura Fernandes, thank you also for putting up with me, and for helping me to grow through our many philosophical and spiritual discussions.

To my parents, Dr. James and Mrs. Janet Foster, and Jenna and Jeff Spielmann, my sister and brother-in-law, thank you so much for all of your support throughout my life, especially these past seven years. You guys are amazing, and I love you all very much.

Finally, to the love of my life, Grace, thank you so much for all you have put up with these past seven years. You have been incredibly loving, patient and supportive, and have inspired me to become a better man. I definitely could not have done any of this without you. I can't wait to spend the rest of my life with you!

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER 1

Introduction

In recent years, there has been a great deal of interest in using a patient's specific information to make more informed treatment decisions. This practice, often referred to as personalized medicine, is motived by the fact that few treatments are equally effective for all individuals in a population. A treatment may be quite effective for some subset of a population, but mildly effective, ineffective, or even harmful for others. Because of this, there is great interest in identifying which individuals in a population, if any, will respond well to a treatment. More specifically, it is necessary to identify which characteristics, if any, lead to this enhanced response if one wishes to pursue a personalized treatment.

A generally accepted approach to identifying such characteristics is to pre-define a small number of subregions of the covariate space before looking at the data, and then evaluate them. However, one may not always know which subgroups to consider, and increasing the number of subgroups considered also increases the risk of false positive findings, which is already a well-known danger in subgroup analysis. One way to reduce this risk of false positives is to employ a multiple testing approach, such as a Bonferroni correction. This can effectively reduce false positives, but also decreases the power to detect true subgroups.

An alternative approach, which will be the main focus of this dissertation, is to pre-define a statistical procedure for identifying subgroups [Ruberg et al., 2010]. These procedures can eliminate the need for *a priori* specification of subgroups, and may help to reduce the risk of false findings, though some risk still remains. Such procedures have been proposed by many authors, including Friedman and Fisher [1999], Negassa et al. [2005], Su et al. [2008, 2009], Brinkley et al. [2010], Cai et al. [2011], Foster et al. [2011], Lipkovich et al. [2011], Qian and Murphy [2011], Zhao et al. [2011], Imai and Ratkovic [2012] and Zhang et al. [2012]. When developing such a procedure, it is common to first consider what each subject's potential outcome value would be, given each of the treatment options [Zhang et al., 2012]. That is, when two treatment options (say treatment 1 and treatment 0) are present, it is common to consider what each subject's potential outcome would be given treatment 1 and what it would be given treatment 0. One general way to identify subgroups is to first estimate these two potential outcomes, and then take the difference, which is an estimate of the treatment effect. One can then investigate the relationship between these estimated treatment effects and the covariates to obtain subgroups. Alternatively, one could attempt to identify subgroups by maximizing the expected response under a class of treatment regimes [Gunter et al., 2007, Zhang et al., 2012]. The treatment regimes in such a case will generally be defined based on the covariates, and will be of the form $\boldsymbol{x}_i \in A \Rightarrow T_i = 1$, $\boldsymbol{x}_i \notin A \Rightarrow T_i = 0$, where $\boldsymbol{x}_i$ and T_i are the covariate vector and treatment indicator for subject i and A is some region of the covariate space. Thus, with this approach, the "optimal" subgroup is the region A which maximizes the expected response when only individuals in A will receive treatment 1 and only individuals in region A^c will receive treatment 0.

Note that, in the setting of personalized medicine, the ultimate goal is to use the identified subregion to define "rules", which can be used in the future to make more informed treatment decisions. Moreover, these treatment decisions will nearly always be made by someone such as a physician or nurse practitioner, who may only have a modest understanding of statistics. Thus, an issue which should be considered before employing any subgroup identification procedure is the potential form and complexity of the resulting subgroup(s), which can both vary considerably depending on which approach is taken. A very complex subgroup, which depends on some, or perhaps all of the available covariates will often be quite good at identifying truly enhanced responders, but such subgroups may lack "nice" interpretability. In addition, the dependence on a large number of covariates means a large amount of information will need to be collected before a treatment decision can be made, which could lead to slower, more expensive, or more invasive (due to performing of unnecessary procedures to collect information) treatment decisions than are necessary. This could potentially limit the chances of such a subgroup being used in practice. In contrast, a very simple subgroup, perhaps depending on only one or two covariates, may less accurately identify enhanced responders, but will be easier to interpret, and will likely see more real-world use. Given this tradeoff between classification accuracy and interpretability, the most "ideal" subgroups may be those which are of only moderate complexity. To this end, we will focus on the identification of simple subgroups, which depend on only a few covariates, and have a very simple form. In many cases, such subgroups may be able to classify enhanced responders well, while still being nicely interpretable, and will thus be more likely to see real-world use.

Because a large number of covariates may often exist, identifying "useful" subgroups which depend on only a few covariates will generally require some form of variable selection. One simple way to do this is to use a regression tree, which splits the data into a number of regions of the covariate space (i.e. subgroups) [Negassa et al., 2005, Su et al., 2008, 2009, Foster et al., 2011, Lipkovich et al., 2011]. These regions contain individuals who are similar with regard to the response, and they are generally defined using only a subset of the available covariates. Alternatively, one could consider a more model-based approach to selecting covariates. In particular, one could consider some form of penalized regression. Potential penalty functions include the LASSO [Tibshirani, 1996], smoothly clipped absolute deviation (SCAD) [Fan and Li, 2001] and adaptive LASSO [Zou, 2006] penalties. These penalty functions are designed to force the regression parameter estimates which correspond to "useless" variables to zero, thereby removing these variables from the model. Thus, predicted values of the response of interest will be a function of a linear combination of a potentially small number of covariates.

As previously mentioned, most subgroup analysis carries with it some potential for false positive findings. Pre-defining a subgroup identification approach can help to reduce false positives, but in our experience, these methods also have a tendency to identify subgroups, even when no true enhanced subgroup exists. To further reduce the potential for false findings, one may wish to evaluate the "usefulness" of a subgroup once it is identified, perhaps by performing a hypothesis test or computing some type of "enhancement" metric [Foster et al., 2011]. Before such a hypothesis test can be implemented, one must consider what "null" means in the setting of subgroup analysis. Note that "meaningful" subgroups arise when the treatment effect differs as covariate values differ. That is, subgroups arise when treatment-by-covariate interactions exist. Thus, in this setting, one could define "null" data as that in which treatment is constant with respect to the covariates, so that no treatment-by-covariate interactions exist. Alternatively, if one has a specific subgroup to evaluate, “null” data could be defined as data for which the effect of treatment in the chosen subgroup is no different than that for the entire population. In this dissertation, we will consider the more general definitition that no treatment-by-covariate interactions exist.

In this dissertation, we will consider a number of methods which use randomized clinical trial data to identify simple subgroups of enhanced treatment effect, which should depend on only a small number of covariates. Using randomized clinical trial data allows us to avoid potential problems with confounded relationships, and thus have more confidence that the identified subgroup contains “truly” enhanced individuals. Moreover, randomized clinical trial data generally contains a large number of subjects, which is advantageous, as large samples are generally required if one wishes to accurately identify subgroups.

In Chapter 2, we consider the use of adaptive LASSO-penalized monotone single-index models to identify subgroups of enhanced treatment effect. A single-index model assumes that the outcome of interest is an unknown function of a linear combination of covariates. By forcing this unknown function to be monotone, we are able to nicely describe the estimated effect of each covariate on the response of interest. In addition, by penalizing the regression parameter, the resulting model will generally depend on only a relatively small number of covariates, making it easier to interpret.

In Chapter 3, we propose a two-stage subgroup identification procedure, which can be viewed as a model-based alternative to Virtual Twins [Foster et al., 2011]. In the first stage of this procedure, we use nonparametric regression to obtain treatment effect estimates for each subject. From these estimates, we define a criterion [Sutton and Barto, 1998, Gunter et al., 2007], which is subsequently used to systematically evaluate many subgropus of a simple, pre-specified form. The identified subgroup is that which has the best value of the evaluation criterion. In this chapter, we also consider the use of an enhancement metric [Foster et al., 2011] to evaluate the utility of identified subgroups.

In Chapter 4, we propose a number of permutation-based methods for obtaining p-values for treatment-by-covariate interactions, which can be used to test whether or not an identified subgroup is truly enhanced. These methods are used to obtain p-values for some of the enhancement metric estimates discussed in Chapter 3. All methods in this dissertation are evaluated in simulation studies, and illustrated using randomized clinical trial data.

In Chapter 5, we present an overall discussion of the proposed methods, and consider a number of potential extensions and modifications of these methods.

CHAPTER 2

Variable selection in monotone single-index models via the adaptive LASSO

2.1 Introduction

Linear regression is a simple and commonly-used technique for assessing relationships of the form $y = \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon$ between an outcome of interest, y, and a set of covariates, $x_1, \ldots, x_p$; however, in many cases, a more general model may be desirable. As noted by Hardle et al. [1993], one particularly useful and more general variation of the linear regression formulation is the single-index model

$$y_i = \eta(\boldsymbol{\beta}^T \boldsymbol{x}_i) + \epsilon_i, \tag{2.1}$$

where $\boldsymbol{x}_i$'s are subject-specific covariate vectors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, $y_i \in \mathbb{R}$, η is an unknown function, $\epsilon_1, \ldots, \epsilon_n$ are iid errors with mean zero and variance σ^2, and ϵ_i's and $\boldsymbol{x}_i$'s are independent. To ensure identifiability, no intercept is included, and β_1 is assumed to be equal to 1. These models are able to capture important features in high-dimensional data, while avoiding the difficulties associated with high-dimensionality, as dimensionality is reduced from many covariates to a univariate index Yu and Ruppert [2002]. Single-index models have applications to a number of fields, including discrete choice analysis in econometrics and dose-response models in biometrics Hardle et al. [1993].

There is a rich literature on estimation of $\boldsymbol{\beta}$ and η, including Hardle et al. [1993], Yu and Ruppert [2002], Ichimura [1993], Carroll et al. [1997], Xia et al. [2002], Xia and Hrdle [2006], among many others. Additionally, variable selection for single-index models was considered by Kong and Xia [2007], who proposed the separated cross-validation method, and Liang et al. [2010], who applied the smoothly clipped absolute deviation (SCAD) approach to partially linear single-index models. However, little consideration has been given to such problems for monotone single-index models, where η is required to be non-decreasing (or non-increasing). In the case of linear models, a great many authors, including Tibshirani [1996], Fan and Li [2001], Zou [2006] and Zou and Zhang [2009] have considered variable selection via penalized least squares, which allows for simultaneous selection of variables and estimation of regression parameters. Several penalty functions, including the SCAD Fan and Li [2001], the adaptive LASSO Zou [2006] and the adaptive elastic-net Zou and Zhang [2009], have been shown to possess favorable theoretical properties, including the *oracle* properties; that is, consistency of selection and asymptotic normality, with the asymptotic covariance matrix being the same as that which would be obtained if the true underlying model were known. Hence, for large samples, oracle procedures perform as well as if the true underlying model were known in advance. Furthermore, Liang et al. [2010] established the oracle properties for the SCAD for partially linear single-index models. Given the desirable properties of the SCAD, adaptive LASSO and adaptive elastic-net approaches, it is natural to consider the extension of these methods to monotone single-index models. Unlike the adaptive LASSO and adaptive elastic-net, which present a convex optimization problem, the SCAD optimization problem is non-convex, and thus more computationally demanding Hastie et al. [2009]. In addition, the adaptive elastic-net and SCAD methods require the selection of two tuning parameters, whereas the adaptive LASSO requires the selection of a single tuning parameter. Therefore, for convenience, computational efficiency, and because covariates in our example data are not highly correlated (a condition under which the adaptive elastic-net is especially good), we consider adaptive LASSO penalized least squares estimation of $\boldsymbol{\beta}$ in monotone single-index models.

The assumption of monotonicity and the desire to select a subset of the covariates are motivated in part by the randomized clinical trial data considered in Foster et al. [2011]. A monotonicity assumption is often reasonable, and such an assumption may improve prediction and reduction in model complexity, while also allowing for more straightforward inference. Foster et al. Foster et al. [2011] consider methods for subgroup identification in randomized clinical trial data. In such cases, should a subgroup be identified, it is desirable that this subgroup be easily described, and depend on only a small number of covariates. Application of the methods proposed in this paper result in estimates $\hat{\eta}$ and $\hat{\boldsymbol{\beta}}$, such that $\hat{y}_i = \hat{\eta}(\hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)$, where $\hat{\eta}$ is monotone and $\hat{\boldsymbol{\beta}}$ generally includes a number of zero values. Using this model, one can classify individuals with $\hat{y}$'s beyond some predefined threshold, c, as being in the subgroup. Then, because of the monotonicity of $\hat{\eta}$, the predefined threshold can be converted into an equivalent threshold, c', on $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$, and the impact of the chosen covariates on subgroup membership can be easily described. Without the assumption of monotonicity, the subgroup may be a collection of several disjoint subregions of the covariate space, making each covariate's impact on subgroup membership more difficult to ascertain.

The remaining sections of this article are as follows. In Section 2, we consider penalized least-squares estimation for monotone single-index models, briefly discuss asymptotics, and discuss a method to obtain standard error estimates for $\boldsymbol{\beta}$. In Section 3, we present the results of a simulation study implemented to assess the performance of the adaptive LASSO penalized single-index models. In Section 4, we briefly discuss the application of this method to the randomized clinical trial data, and in Section 5, we give concluding remarks.

2.2 Estimation for monotone single index models

Our estimation procedure iterates between estimation of $\boldsymbol{\beta}$ and η until convergence. Given some η, the penalized least-squares estimator of $\boldsymbol{\beta}$ can be found by minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big(y_i - \eta(\boldsymbol{\beta}^T \boldsymbol{x}_i)\big)^2 + \lambda_n \sum_{j=2}^{p} w_j |\beta_j|, \tag{2.2}$$

where $w_j,\ j = 2, \dots, p$ are known weights and covariates $\boldsymbol{x}_i$ are standardized to have mean zero and variance 1. Due to the identifiability constraint specified in model (2.1), β_1 is not penalized. Following Zou [2006], we choose $w_j = |\hat{\beta}_{init,j}|^{-\gamma}$ for $\gamma > 0$, where $\hat{\boldsymbol{\beta}}_{init}$ is a n^{α}-consistent estimator of $\boldsymbol{\beta}$, where $0 < \alpha \leq \frac{1}{2}$. We use linear ordinary least squares (OLS) estimates for $\hat{\boldsymbol{\beta}}_{init}$, as under the assumptions of Theorem 2.1 in Li and Duan [1989], these are shown to be $\sqrt{n}$-consistent up to a multiplicative scalar. Once obtained, $\hat{\boldsymbol{\beta}}_{init}$ is rescaled by $\hat{\beta}_{1,init}$. Alternatively, weights could be defined using the unpenalized single-index model estimates of $\boldsymbol{\beta}$.

For a given $\boldsymbol{\beta}$, without considering the monotonicity constraint, η can be estimated at some point t using the Nadaraya-Watson kernel-weighted average:

$$\hat{\eta}(t; \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}, h) = \frac{\sum_j y_j K\left(\frac{t-\boldsymbol{\beta}^T\boldsymbol{x}_j}{h}\right)}{\sum_j K\left(\frac{t-\boldsymbol{\beta}^T\boldsymbol{x}_j}{h}\right)}, \tag{2.3}$$

where $\boldsymbol{X}$ is the covariate matrix, K is a fixed kernel function and h is a bandwidth. Note that, when $\boldsymbol{\beta}$ is known, $\hat{\eta}$ is determined by h, so a value of h must be chosen. We consider kernel functions which are symmetric probability densities. For numerical stability, we hold (2.3) fixed for all t outside the range of the $\boldsymbol{\beta}^T\boldsymbol{x}$'s. That is, $\hat{\eta}(t) = \hat{\eta}(\min_i(\boldsymbol{\beta}^T\boldsymbol{x}_i))$ if $t < \min_i(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ and $\hat{\eta}(\max_i(\boldsymbol{\beta}^T\boldsymbol{x}_i))$ if $t > \max_i(\boldsymbol{\beta}^T\boldsymbol{x}_i)$.

Combining (2.2) and (2.3), the adaptive LASSO estimator for $\boldsymbol{\beta}$ is obtained by minimizing

$$\hat{Q}(\boldsymbol{\beta}, h) = \sum_i{}' \left(y_i - \hat{\eta}(\boldsymbol{\beta}^T\boldsymbol{x}_i; \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{X}, h)\right)^2 + \lambda_n \sum_{j=2}^{p} w_j|\beta_j| \tag{2.4}$$

with respect to $\boldsymbol{\beta}$, where $\sum_i'$ denotes summation over i such that the denominator in the kernel estimator is not too close to zero. Details can be found in Hardle et al. [1993]. With the inclusion of the penalty term in (2.4), $\hat{\boldsymbol{\beta}}$ becomes a function of λ_n, so in addition to h, a value of λ_n must be chosen if $\hat{\boldsymbol{\beta}}$ is to be obtained. Throughout this paper, we refer to the method of estimating $\boldsymbol{\beta}$ and η without a monotonicity constraint, using objective function (2.4), as the *unconstrained* approach.

2.2.1 A smooth monotone function estimate for η with fixed $\boldsymbol{\beta}$

There are a variety of ways to obtain smooth monotone regression function estimates, including quadratic B-splines He and Shi [1998], I-splines Ramsay [1988], empirical distribution tilting Hall and Huang [2001], the scatterplot smoothing approach of Friedman and Tibshirani [1984], and the kernel-based approach of Mukerjee [1988] and Mammen [1991]. We consider the kernel-based method of the latter two papers, which we briefly describe below.

Assume $\boldsymbol{\beta}$ is known. The proposed monotone estimator $\hat{\eta}_m$ requires two steps:

Isotonization. This step involves the application of the pooled adjacent violator algorithm (PAVA) Barlow et al. [1972]. Using $(\boldsymbol{\beta}^T \boldsymbol{x}_i, y_i)$ ordered by increasing $\boldsymbol{\beta}^T \boldsymbol{x}_i$ as data, PAVA produces monotone estimates $\hat{m}_1, \ldots, \hat{m}_n$, which are averages of y_j's near i (unless y's are already monotone, in which case $\hat{m}_i = y_i$), and which are not necessarily smooth Friedman and Tibshirani [1984].

Smoothing. Apply the kernel estimator (2.3) with y_i replaced by $\hat{m}_i$ for all i to estimate η. That is, $\hat{\eta}_m(t) = \hat{\eta}(t; \boldsymbol{\beta}, \widehat{\boldsymbol{m}}, \boldsymbol{X}, h)$.

Since $\hat{m}_1, \ldots, \hat{m}_n$ are monotone, the resulting function estimate is monotone in t. It is worth noting that this may not necessarily be the case for other smoothing methods, such as local linear regression.

As previously mentioned, a bandwidth is needed to estimate η, and can be found using cross-validation; however, our algorithm requires estimation of both η and its derivative η', so care must be taken. In particular, to ensure good algorithmic convergence, it is crucial that $\hat{\eta}'$ be smooth, but to obtain a smooth estimate of η', it is often necessary to oversmooth η. Thus, we restrict the range of potential bandwidths in our cross-validation. Specifically, h is restricted to be between $0.1{*}sd(\boldsymbol{X}\boldsymbol{\beta})$ and $sd(\boldsymbol{X}\boldsymbol{\beta})$, as values in this range were found to perform well in our simulations.

2.2.2 Estimation for $\boldsymbol{\beta}$ with fixed η

The shooting algorithm proposed by Fu [1998] has been shown to perform well in solving LASSO penalized least-squares problems for linear models Friedman et al. [2007]. Therefore, we consider the application of this algorithm to LASSO problems for the single-index model. One way to achieve this is to employ a linear approximation via Taylor series expansion of $\eta(\boldsymbol{\beta}^T \boldsymbol{x}_i)$ about $\boldsymbol{\beta}_0^T \boldsymbol{x}_i$, where $\boldsymbol{\beta}_0$ is known. We define the linear approximation as follows:

$$\eta\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right) \approx \eta\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right) + \eta^{'}\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right)\left[\boldsymbol{\beta}^T \boldsymbol{x}_i - \boldsymbol{\beta}_0^T \boldsymbol{x}_i\right]. \tag{2.5}$$

Let

$$y_i^* = y_i - \eta\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right) + \eta^{'}\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right)\boldsymbol{\beta}_0^T \boldsymbol{x}_i$$

and

$$\boldsymbol{x}_i^* = \eta^{'}\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right)\boldsymbol{x}_i.$$

Then we have

$$y_i - \eta\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right) \approx y_i - \eta\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right) - \eta^{'}\left(\boldsymbol{\beta}_0^T \boldsymbol{x}_i\right)\left[\boldsymbol{\beta}^T \boldsymbol{x}_i - \boldsymbol{\beta}_0^T \boldsymbol{x}_i\right] = y_i^* - \boldsymbol{\beta}^T \boldsymbol{x}_i^*,$$

and (2.4) can be approximated by

$$\hat{Q}_{lin}\left(\boldsymbol{\beta}\right) = \sum_i{}^{'}\left(y_i^* - \boldsymbol{\beta}^T \boldsymbol{x}_i^*\right)^2 + \lambda_n \sum_{j=2}^{p} w_j |\beta_j|, \tag{2.6}$$

which is a LASSO penalized least-squares problem for the linear model, and can thus be solved using the shooting algorithm.

Note that (2.5) involves an estimate of η'. This estimate is obtained as follows. Sort the observations by increasing $\boldsymbol{\beta}_0^T \boldsymbol{x}_i$, and define new data $\Big\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) \ : \ , \ i = 1, \ldots, n-1\Big\}$, where $\tilde{y}_i = \frac{\eta(\boldsymbol{\beta}_0^T \boldsymbol{x}_{i+1}) - \eta(\boldsymbol{\beta}_0^T \boldsymbol{x}_i)}{\boldsymbol{\beta}_0^T \boldsymbol{x}_{i+1} - \boldsymbol{\beta}_0^T \boldsymbol{x}_i}$, and $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \frac{\boldsymbol{x}_{i+1} - \boldsymbol{x}_i}{2}$. This new data should "look like" data coming from the model $\tilde{y}_i = \eta'(\boldsymbol{\beta}^T \boldsymbol{x}_i) + \tilde{\epsilon}_i$, so $\eta'(t)$ can be estimated using (2.3), but with $\Big\{(\boldsymbol{x}_i, y_i) \ : \ i = 1, \ldots, n\Big\}$ replaced by $\Big\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) : \ i = 1, \ldots, n-1\Big\}$, i.e. $\hat{\eta}'(t) = \hat{\eta}(t; \boldsymbol{\beta}, \tilde{\boldsymbol{y}}, \tilde{\boldsymbol{X}}, \tilde{h})$, where $\tilde{h}$ is a new bandwidth for the derivative estimate. To select $\tilde{h}$, cross-validation can again be used.

2.2.3 Algorithm

The algorithm to obtain final estimates of η and $\boldsymbol{\beta}$ iterates between the steps in Sections 2.1 and 2.2 until convergence. After k iterations, let $\hat{\boldsymbol{\beta}}^{(k)}$, $\hat{\eta}_m^{(k)}$ and $\hat{m}_1^{(k)}, \ldots, \hat{m}_n^{(k)}$ denote the current estimates of $\boldsymbol{\beta}$ and η and the current PAVA estimates respectively. For a given λ_n, the "final" estimates of $\boldsymbol{\beta}$ and η are obtained as follows:

1. Using data $\Big\{(\hat{\boldsymbol{\beta}}^{(k)T} \boldsymbol{x}_i, y_i) : \ i = 1, \ldots, n\Big\}$, apply PAVA to obtain new monotone data $\Big\{(\hat{\boldsymbol{\beta}}^{(k)T} \boldsymbol{x}_i, \hat{m}_i^{(k+1)}) : \ i = 1, \ldots, n\Big\}$, and define the monotone function estimate $\hat{\eta}_m^{(k+1)}(t)$ using (2.3). Select h via a grid search on values $\Big\{0.1^* sd(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})$, $0.2^* sd(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}), \ldots, \ sd(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})\Big\}$ using leave-one-out cross-validation. For computational convenience, fix h after a small number, say a, of iterations (i.e. when $k = a$).

2. Using data $\Big\{(\hat{\boldsymbol{\beta}}^{(k)T} \boldsymbol{x}_i, \hat{m}_i^{(k+1)}) \ : \ i = 1, \ldots, n\Big\}$, obtain the derivative data $\Big\{(\hat{\boldsymbol{\beta}}^{(k)T} \tilde{\boldsymbol{x}}_i, \tilde{y}_i^{(k+1)}) : \ i = 1, \ldots, n-1\Big\}$, and define the derivative $\hat{\eta}^{(k+1)'}(t)$ using

(2.3). Select $\tilde{h}$ from the grid $\Big\{0.1^{*}sd(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}), 0.2^{*}sd(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)}), \ldots, sd(\boldsymbol{X}\hat{\boldsymbol{\beta}}^{(k)})\Big\}$ using leave-one-out cross-validation. As with h, $\tilde{h}$ is fixed after a iterations.

3. Let the general notation $z^{(k^l)}$ indicate the l^{th} update to $z^{(k)}$. Using approximation (2.5), obtain data $\Big\{(\boldsymbol{x}_i^{*(k^1)}, y_i^{*(k^1)}) : \; i = 1, \ldots, n\Big\}$, and minimize $\hat{Q}_{lin}^{(k^1)}(\boldsymbol{\beta})$ from (2.6), giving $\hat{\boldsymbol{\beta}}^{(k^1)}$. Repeat this step $m-1$ more times, for a total of m iterations, each time updating the linear approximation (2.5), so that $\hat{\boldsymbol{\beta}}^{(k^m)} \equiv \hat{\boldsymbol{\beta}}^{(k+1)}$ comes from data $\Big\{(\boldsymbol{x}_i^{*(k^m)}, y_i^{*(k^m)}) : \; i = 1, \ldots, n\Big\} \equiv \Big\{(\boldsymbol{x}_i^{*(k+1)}, y_i^{*(k+1)}) : \; i = 1, \ldots, n\Big\}$.

4. Cycle through steps 1-3 until $\Big\|\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}\Big\|$ becomes smaller than a prespecified precision level. The final estimate of η is then obtained by implementing step 1 once more using the converged $\boldsymbol{\beta}$ estimate.

The identifiability constraint is imposed by rescaling $\hat{\boldsymbol{\beta}}^{(k)}$ by $\hat{\beta}_1^{(k)}$ each time Step 3 is completed, so it is desirable that β_1 be nonzero to avoid potential numerical problems. To help ensure this in practice, one could first fit a linear model, and choose the largest (or most significant) β_j estimate to be that which is subsequently unpenalized and forced to be 1. If in the final model another coefficient is larger, then one could re-run the analysis with that coefficient being the one which is upenalized and forced to be 1. As suggested by one of the reviewers, one could also consider a sensitivity analysis in which multiple models were fit, each time forcing a different coefficient to be 1.

A value of λ_n must be chosen before $\boldsymbol{\beta}$ can be estimated. Suppose that $\hat{\boldsymbol{\beta}}(\lambda_n)$ and $\hat{\eta}_m(t; \lambda_n)$ are the estimates of $\boldsymbol{\beta}$ and $\eta(t)$, given tuning parameter λ_n. To choose a value of λ_n, we use the Bayes information criterion (BIC) measure of Liang et al.

[2010]. Specifically, we choose the value of λ_n that minimizes:

$$BIC(\lambda_n) = \log\Big\{\frac{1}{n}\sum_i{}' \Big(y_i - \hat{\eta}_m(\hat{\boldsymbol{\beta}}(\lambda_n)^T\boldsymbol{x}_i;\lambda_n)\Big)^2\Big\} + \frac{\log(n)}{n}DF_{\lambda_n},$$

where DF_{λ_n} is one less than the number of non-zero values in $\hat{\boldsymbol{\beta}}(\lambda_n)$, since $\hat{\beta}_1$ is forced to be nonzero. To find the optimal λ_n, a grid search is employed.

In the remaining sections, the monotone-constrained method described above is referred to as the *constrained* approach.

2.2.4 Asymptotics

Using the results of Hardle et al. [1993] and arguments similar to Zou [2006], it is possible to establish the *oracle* properties for the unconstrained approach. We provide an outline of such an argument here.

Suppose the regularity conditions of Hardle et al. [1993] hold. Then, by their main theorem, we can rewrite the sum of squares portion of (2.4) as a sum of three terms, $\tilde{S}$, T, and R, where $\tilde{S}$ and T depend only on $\boldsymbol{\beta}$ and h respectively, and the remainder term R is negligible. Thus, as $\tilde{S}$ is the only term which depends on $\boldsymbol{\beta}$, (2.4) can be reduced to $\tilde{S}(\boldsymbol{\beta}) + \lambda_n\sum_{j=2}^{p} w_j|\beta_j| = n\Big\{\boldsymbol{W}_0^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\Big\}^T\Big\{\boldsymbol{W}_0^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\Big\} + \lambda_n\sum_{j=2}^{p} w_j|\beta_j|$, where $\boldsymbol{W}_0$ is a $p\times p$ matrix, $\boldsymbol{\beta}_0$ is the true index parameter, and $\boldsymbol{Z}$ is an asymptotically normal $N(\boldsymbol{0},\boldsymbol{I})$ p-vector. Now suppose that $\frac{\lambda_n}{\sqrt{n}} \to 0$, and $\lambda_n n^{(\gamma-1)/2} \to \infty$, where $\gamma \in (0, \frac{3}{5}]$, and let $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}$, where $\|\ \boldsymbol{u}\|\ \le C$. From here, following arguments very similar to Zou [2006], the *oracle* properties can be established. Two theorems and a formal proof for the unconstrained approach are given in the Appendix.

It seems that the *oracle* properties will also hold for $\boldsymbol{\beta}$ estimates from the constrained approach under certain conditions. Specifically, under the conditions of Theorem 2 in Mammen [1991], we have $\hat{\eta}_m(t) = \hat{\eta}(t) + O_p(n^{-8/15})$, for all t, where $\hat{\eta}_m$ is our monotone estimator of η and $\hat{\eta}$ is the Nadaraya-Watson kernel-weighted average. Thus, it is possible to reduce the penalized sum of squares for the constrained approach to (2.4) plus a negligible remainder term. The *oracle* properties for the constrained approach would hold by the same reasoning used for the unconstrained approach.

In practice it is difficult to verify that the conditions needed for the theory hold. Because a data-driven method (BIC) is used to select the tuning parameter, λ_n, we cannot guarantee the required rate of convergence. Thus, the assumptions $\frac{\lambda_n}{\sqrt{n}} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$ may not hold.

2.2.5 Bootstrap standard errors

Standard errors for our $\boldsymbol{\beta}$ estimates can be obtained via the bootstrap. In particular, for a given data set, we employ the adaptive LASSO-based residual bootstrap (ARB) approach discussed by Chatterjee and Lahiri [2011] to obtain many, say M, bootstrap data sets. A penalized single-index model is then fit on each of these bootstrap data sets, giving M sets of estimates. The estimated standard errors are then obtained by taking the standard deviations of the M estimates for each β_j.

For a given data set, we obtain a residual bootstrap data set as follows. Suppose $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}$ are final estimates of $\boldsymbol{\beta}$ and η for a particular data set. Let $e_i = y_i - \hat{\eta}(\hat{\boldsymbol{\beta}}^T \boldsymbol{x_i})$, $i = 1, \ldots, n$ be the residuals for this data set. A residual bootstrap data set is then obtained by replacing y_i with $\hat{\eta}(\hat{\boldsymbol{\beta}}^T \boldsymbol{x_i}) + e_i^*$, $i = 1, \ldots, n$, where $\{e_1^*, \ldots, e_n^*\}$ is a random sample (drawn with replacement) from the centered residuals, $e_i - \frac{1}{n}\sum_{i=1}^n e_i$, $i = 1, \ldots, n$. The covariate matrix remains the same across the bootstrap data sets.

Based on additional simulations (results not given), creating residual bootstrap data sets using permuted (sampled without replacement) residuals gives nearly identical results to those shown in Table 2.2. As noted by one reviewer, in practice, the interpretation for the standard errors can be awkward, particulary in cases where a number of covariates are highly correlated. In such cases, one might expect the distribution of these estimates to be a mixture of a continuous distribution and a point mass at zero. Thus, the estimates are a product of both selection and estimation, which can make interpretation difficult. This may be due to the known shortcomings of the adaptive LASSO for highly correlated predictors. If one believes that a number of covariates may be highly correlated, an alternative approach, such as the adaptive elastic-net, may perform better, and may lead to bootstrap standard error estimates based on a smaller number of zeros.

We generally suggest that one reselect λ_n with each bootstrap data set; however, our simulations (results not shown) suggest that holding λ_n fixed across bootstrap data sets gives standard error estimates which are nearly identical to those found by reselecting λ_n for each bootstrap data set. Thus, it may be reasonable to consider fixed-λ_n bootstrap standard errors if reselecting λ_n for each bootstrap data set is too computationally burdensome.

2.3 Simulations

A simulation study was performed using R software to evaluate the performance of the proposed methods. To comply with the conditions in Section 2.4, a value of $\frac{3}{5}$ was chosen for γ for adaptive LASSO. Additionally, for each example, a large test set ($n = 10,000$) was generated, and final $\boldsymbol{\beta}$ estimates from each of the simulated data sets were used to calculate the mean squared error (MSE) for this large test set. To evaluate the performance of all methods considered, we recorded the number of correct and incorrect zero values in $\hat{\boldsymbol{\beta}}$, as well as the total proportion of $\hat{\beta}_j$'s correctly estimated as zero or non-zero for each data set. The average of these proportions across all simulated data sets is referred to in Table 2.1 as the *relative frequency correct*. We also computed the false discovery rate (FDR), which is the percentage of non-zero $\hat{\boldsymbol{\beta}}$ values which should have been zero. For each data set, the optimal tuning parameter value λ_n was chosen from the grid $\{0, 0.01, \ldots, 0.25\}$ using BIC.

2.3.1 Examples

For all simulations, 100 data sets of size 100 were generated from the model

$$y_i = (\boldsymbol{\beta}^T \boldsymbol{x}_i)^3 + \epsilon_i,$$

where $\boldsymbol{x}_i$'s were $Unif[-\frac{1}{2}, \frac{1}{2}]$, and error terms were normal with mean zero and variance σ^2. We considered five different cases:

(i) $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0)^T$, $\boldsymbol{x}_i$'s independent, and ϵ's independent with $\sigma = 0.20$;

(ii) Same as case (i), but with $\sigma = 0.30$;

(iii) Same as case (i), but with $\boldsymbol{\beta}$ changed to $(1, 0.8, 0, 0, 0, 0, -0.2, 0, 0, 0)^T$;

(iv) Same as case (i), but with $Corr(x_{ij}, x_{ik}) = 0.5, \ j \neq k$;

(v) Same as case (i), but with an additional 50 noise covariates, so that $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0, \mathbf{0}_{1\times 50})^T$.

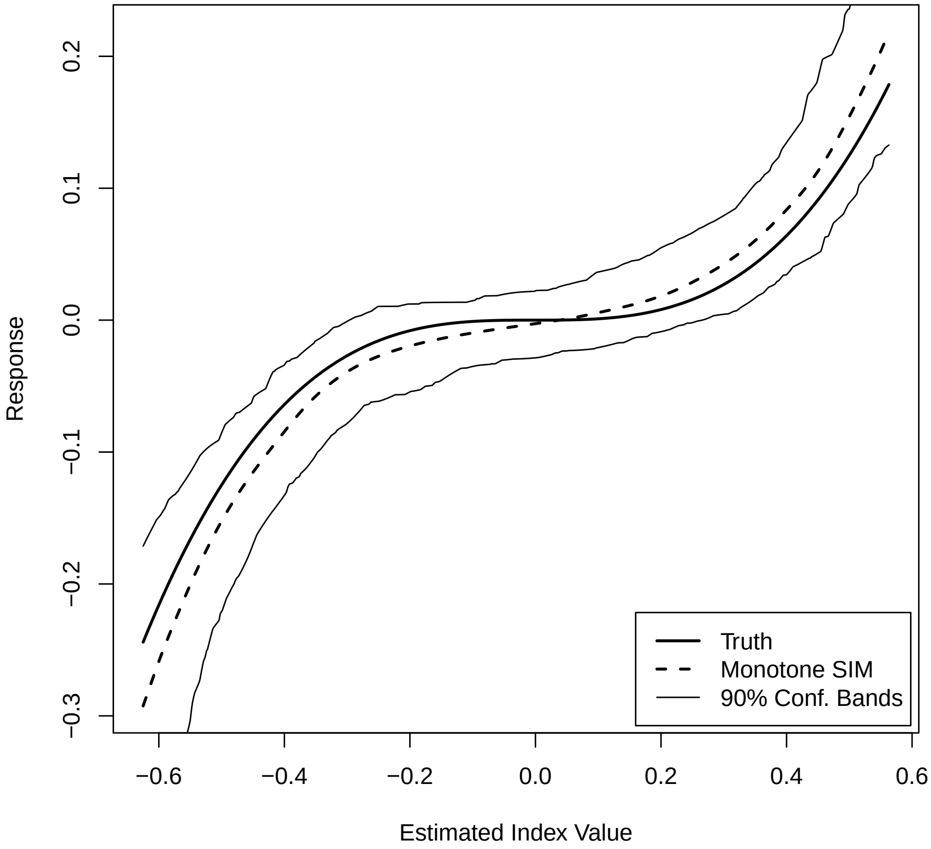


Figure 2.1: Average $\hat{\eta}_m$ values from 100 simulations

From Table 2.1, we can see that, in all cases, the constrained approach shows noticeably better reduction in model complexity and smaller FDR than the unconstrained approach. Additionally, the constrained approach has mean test MSEs which are smaller, and closer to the corresponding oracle test MSEs than the unconstrained approach. Reduction in model complexity for the constrained approach appears to be reasonably insensitive to the changes in simulation settings considered above; however, the unconstrained approach appears to suffer in this regard, especially when true parameter values are decreased or error standard deviation is increased.

Table 2.1: Simulation results: variable selection performance

Method	Rel. Freq Correct	Avg. No. $\hat{\boldsymbol{\beta}} = 0^*$ Correct	Avg. No. $\hat{\boldsymbol{\beta}} = 0^*$ Incorrect	FDR	Mean Test MSE ($\times$100)
Case (i)[1]					
Cons.:	0.92	6.22	0.00	0.21	4.93
Uncons.:	0.85	5.47	0.01	0.34	5.63
Cons. oracle:	1.00	7.00	0.00	0.00	4.74
Uncons. oracle:	1.00	7.00	0.00	0.00	4.96
Case (ii)[2] **					
Cons.:	0.89	6.09	0.17	0.24	10.87
Uncons.:	0.75	4.68	0.16	0.45	12.84
Cons. oracle:	1.00	7.00	0.00	0.00	10.07
Uncons. oracle:	1.00	7.00	0.00	0.00	10.56
Case (iii)[3]					
Cons.:	0.86	6.25	0.65	0.24	4.64
Uncons.:	0.73	4.83	0.56	0.47	5.51
Cons. oracle:	1.00	7.00	0.00	0.00	4.45
Uncons. oracle:	1.00	7.00	0.00	0.00	4.65
Case (iv)[4]					
Cons.:	0.88	6.15	0.32	0.24	4.79
Uncons.:	0.82	5.45	0.27	0.36	5.38
Cons. oracle:	1.00	7.00	0.00	0.00	4.51
Uncons. oracle:	1.00	7.00	0.00	0.00	4.78
Case (v)[5]					
Cons.:	0.94	53.70	0.10	0.53	5.64
Uncons.:	0.87	49.36	0.13	0.74	7.16
Cons. oracle:	1.00	57.00	0.00	0.00	4.74
Uncons. oracle:	1.00	57.00	0.00	0.00	4.96

Note: "oracle" indicates true zero $\boldsymbol{\beta}$ values known. η is estimated in all methods. * Average number of variables dropped in final model.

[1] $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0)^T$, $Corr(x_{ij}, x_{ik}) = 0$, $j \neq k$, $\sigma = 0.20$.

[2] Same as Case (i), but $\sigma = 0.3$. [3] Same as Case (i), but $\beta_7 = -0.2$.

[4] Same as Case (i), but $Corr(x_{ij}, x_{ik}) = 0.5$, $j \neq k$.

[5] Same as Case (i), but $\boldsymbol{\beta} = (1, 0.8, 0, 0, 0, 0, -0.7, 0, 0, 0, \mathbf{0}_{1\times 50})^T$.

** Required 101 simulated data sets due to numerical problems.

Additional simulations were implemented to evaluate the performance of the proposed methods under alternative monotonic functions, η (results not shown). In particular, we considered a linear function, and two spline functions; one resembling the cubic function from the above examples, but with two knots chosen to create a wider "flat" section around the origin, and one which is constant to the left of the origin and quadratic to the right. As expected, both methods performed well in the linear case. In the cubic spline case, reduction in model complexity was good, but mean test MSE became noticeably larger, and in the case of the constant spline with the quadratic knot, mean test MSE was good, but reduction in model complexity was noticeably worse. Thus, as one might expect, the proposed methods are less useful in cases where η contains large sections which are nearly flat, or exactly constant.

To evaluate the performance of our standard error estimates, residual bootstrap standard errors (based on 100 bootstrap data sets) were calculated for case (i) above. Let SD denote the standard deviation of the 100 β_j estimates, $j = 1, \ldots, p$. Additionally, let SE and SE$_{sd}$ denote the mean and the standard deviation of the 100 estimated SE's respectively. Looking at Table 2.2, we can see that the standard error estimates appear to perform reasonably well, though they sometimes slightly underestimate or overestimate the true values.

To demonstrate the ability of penalized monotone single-index models to capture non linear relationships, we computed $\hat{\eta}_m$ values across a fine grid of input values and averaged these $\hat{\eta}_m$'s across the 100 data sets in case (i). These average values can be found in Figure 2.1, along with the true function η and 90% empirical pointwise confidence bands for $\hat{\eta}_m$. As we can see, the monotone function estimate $\hat{\eta}_m$ appears to closely follow the true function.

Table 2.2: Performance of standard error estimates

	$\hat{\beta}_2$		$\hat{\beta}_7$	
Method	SD	SE (SE$_{sd}$)	SD	SE (SE$_{sd}$)
Cons.:	0.25	0.20 (0.08)	0.20	0.18 (0.06)
Uncons.:	0.28	0.34 (0.41)	0.28	0.27 (0.26)

Note: required 102 simulated data sets due to numerical problems.

Because we are interested in using the proposed methods to identify subgroups, we also compared "enhancement" classification between the two methods for case (i). For this comparison, we consider a subject to be enhanced if $\eta(\boldsymbol{x}^T\boldsymbol{\beta}) > 0$. On average, 88% of subjects identified as enhanced by the constrained approach were truly enhanced, whereas for the unconstrained approach, only 73% were correctly identified on average. Thus, the constrained approach may be advantageous for applications to subgroup identification.

The two methods methods require approximately the same amount of time to complete a single iteration of our algorithm for a given value of λ. However, for some data sets, the constrained approach requires more iterations to achieve the same degree of convergence as the unconstrained approach. For example, for case (i) of our simulations, the median run time for a data set for the constrained approach was approximately 64% longer than that for the unconstrained approach.

2.4 Example data

In this example, we apply the proposed methods to the Eli Lilly data in Foster et al. [2011], which come from a randomized, double-blinded clinical trial in patients with a critical illness in the ICU conducted over a decade ago. We consider 1019 individuals, of whom 512 received the experimental treatment in addition to the standard of care. The remaining patients received placebo with the standard of care. The intervention is a drug that is intended to improve survival in patients with a critical illness, and the endpoint was survival at 28 days post-randomization to treatment/placebo. We consider 58 covariates analyzed by Foster et al. [2011], which include demographic, laboratory, medical history and questionnaire data. Of these, 9 are binary, 22 are regarded as continuous, and 27 are dummy variables coming from subdivision of 12 categorical variables.

In Foster et al. [2011], a random forest was used to obtain two predicted probabilities, $\hat{P}_{1i}$ and $\hat{P}_{0i}$, for each individual, where P_{1i} is the probability of survival at 28 days post-randomization for subject i if that individual had received treatment and P_{0i} is that if subject i had received placebo. The estimation of these probabilities was motivated by the fact that the methods of Foster et al. [2011] were designed to identify subgroups of enhanced treatment effect in randomized clinical trial data. Therefore, a new outcome representing the treatment effect for person i, $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$, $i = 1, \ldots, n$, was subsequently defined, since individuals in such a subgroup should ideally have values of P_{1i} which are much larger than P_{0i}. Then, a single regression tree was fit using Z as the outcome and the covariates as predictors. This tree identified subgroups of enhanced treatment effect which depended on age at admission, baseline creatinine clearance, baseline interleukin 6 and hypertension (yes, no or unknown). This method was referred to by Foster et al. [2011] as “Virtual Twins.”

Using Z as the outcome and the 58 covariates as predictors, we fit penalized single-index models with and without monotonicity constraints. All covariates were

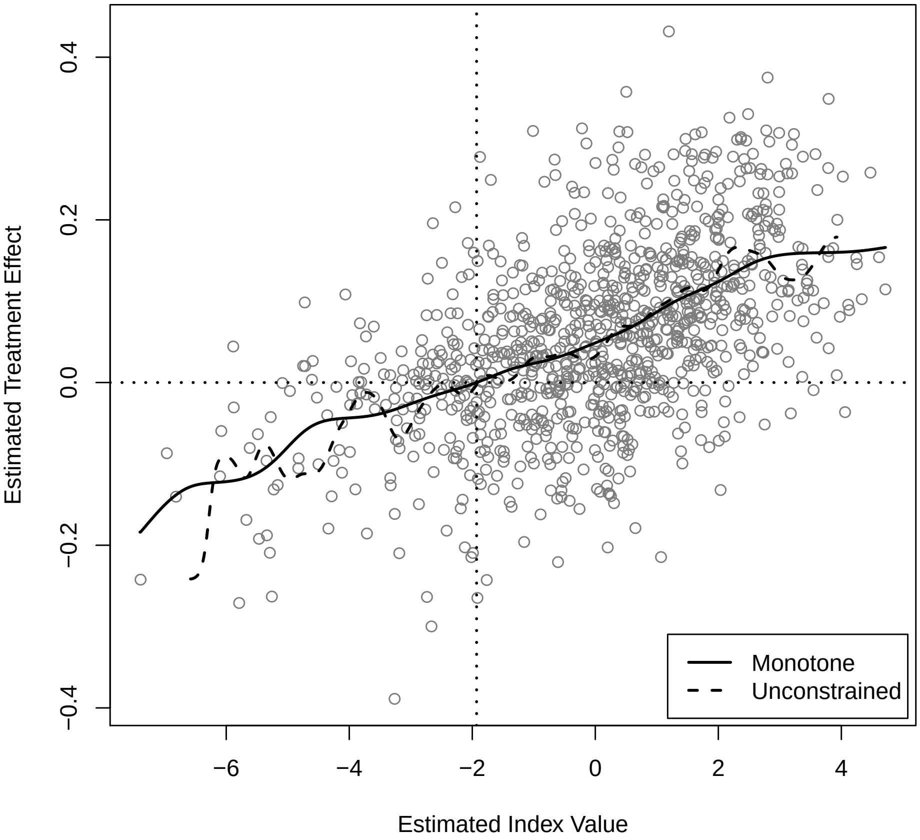


Figure 2.2: Estimates of function $\hat{\eta}(\cdot)$ from Eli Lilly data. Index values in the plotted data are $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$, where $\hat{\boldsymbol{\beta}}$ comes from the constrained approach, and treatment effect estimates are the Z values from Virtual Twins procedure. Those points to the right of the vertical dotted line would be considered "enhanced" based on this analysis.

standardized in this analysis due to large differences in scale, and age at admission was chosen to be the first column of $\boldsymbol{X}$, as its corresponding initial estimate was the largest and most significant value of $\hat{\boldsymbol{\beta}}_{init}$ It should be noted that this analysis was also performed with baseline creatinine clearance as the first column (results now shown), and the same six additional covariates were chosen, along with one other. The relative magnitude of the coefficients in this analysis were similar for most variables. Results from these models (with age at admission as first column of $\boldsymbol{X}$) can be found in Table

2.3. Estimates for the constrained and unconstrained approaches were fairly similar, though an additional covariate, baseline index of independence in activities of daily living (ADL) Katz and Akpom [1976], was included by the constrained approach.

Table 2.3: Estimates for Eli Lilly data

	Unconstrained		Constrained	
Variable	Estimate	SE	Estimate	SE
Age	1.00	-	1.00	-
ADL[1]	-	-	-0.13	0.04
Platelet Count	-0.12	0.03	-0.19	0.08
Creat. Clear.	-0.70	0.17	-0.81	0.21
Interleukin 6	0.60	0.11	0.70	0.13
# Organ Fail.	0.14	0.06	0.23	0.11
APACHE II[2]	0.24	0.09	0.33	0.14

[1] Baseline index of independence in activities of daily living.
[2] Pre-infusion acute physiology and chronic health evaluation II score.

In addition to $\boldsymbol{\beta}$ estimates, we computed bootstrap standard errors using 300 bootstrap samples. Because less important covariates will tend to be removed from the model in most bootstrap samples, resulting in many zero bootstrap estimates, we expect such covariates to have very small bootstrap standard errors.

The six covariates selected by both methods were age at admission, baseline central lab platelet count, baseline creatinine clearance, baseline interleukin 6 (log scale), number of baseline organ failures, and pre-infusion acute physiology and chronic health evaluation II (APACHE II) score, of which age at admission, creatinine clearance and interleukin 6 were also selected by the Virtual Twins method. Plots of the data (from the constrained approach) and the final η estimates can be found in Figure 2.2. We can see that both estimates of η are reasonably close, with the constrained estimate being noticeably more smooth. From Figure 2.2, we can see that the predicted region of enhanced treatment effect consists of $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ values which are larger than approximately -2, with the degree of enhancement increasing as $\hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ becomes larger. The constrained and unconstrained approaches identified 847 and 864 subjects as being enhanced, respectively, and of the 864 identified by the unconstrained approach, 845 were also identified by the constrained approach. Furthermore, for the constrained model, older individuals and those with higher baseline IL-6 respond very well to treatment, and patients with lower baseline creatinine clearance show a greater treatment differential. The findings from this analysis are reasonably consistent with the original conclusions from this trial, which suggested that patients who had higher risk factors for mortality responded better to the treatment.

As both fits suggest a relationship which is close to linear, an adaptive LASSO penalized linear model was also fit (results not shown), once using the default tuning parameter selection settings (10-fold cross-validation using squared error loss) in the R *glmnet* package, and once using BIC to select the tuning parameter. The model resulting from the default tuning parameter selection settings contained 24 covariates, while the model selected using BIC contained 7 covariates. Although BIC is known to give smaller models than cross-validation, this dramatic difference in model complexity was mildly surprising to us. Based on the results of the linear model (using BIC), it appears that the single-index models may not have added much compared to a linear model in this case.

2.5 Discussion

We proposed the use of adaptive LASSO variable selection for monotone single-index models, and showed that it performs well in a variety of situations. The constrained approach noticeably outperformed the unconstrained, and has the advantage of more straightforward interpretation. A linear approximation to η via Taylor series was also proposed, thus allowing for the use of standard LASSO algorithms, such as coordinate descent, which have been shown to perform well. In addition, we suggested the use of residual bootstrap standard errors for $\boldsymbol{\beta}$ estimates, and showed that they perform reasonably well in simulations.

We argue that the unconstrained adaptive LASSO penalized single-index model estimates possess the *oracle* properties when η is estimated using the Nadaraya-Watson formula. Additionally, we briefly argue that, following the results of Mammen [1991], the *oracle* properties may also hold for the constrained approach, and it would be interesting to investigate this more formally. Furthermore, the proof outlined in Section 2.4 assumes that $\boldsymbol{\beta}$ is in a $\sqrt{n}$-neighborhood of the true value, which is likely true given that the initial estimator of $\boldsymbol{\beta}$ is in a $\sqrt{n}$-neighborhood of $\boldsymbol{\beta}_0$.

Our method of obtaining a monotone function estimate is very similar to that of Friedman and Tibshirani [1984]. They suggested that it may be possible to improve the estimation of the monotone penalized single-index model if one considers "one-step" monotone function estimates, such as those suggested by He and Shi [1998] and Ramsay [1988]. This is worthy of further investigation.

The adaptive LASSO penalty was chosen for convenience; however, one may wish to consider other penalty functions. For instance, as noted by a reviewer, the adaptive elastic-net can often outperform the adaptive LASSO approach, particularly when covariates are highly correlated. Note that the linear approximation to the function η does not involve the penalty function. Thus, the proposed method and algorithm could easily be modified if one wished to use a different penalty function, such as the SCAD or adaptive elastic-net.

CHAPTER 3

Selection of simple rules for treatment assignment using patient information

3.1 Introduction

Though some treatments may be more widely effective than others, few, if any, work for all individuals in a target population. In many cases, a treatment may be extremely effective for some subset of a population, but mildly effective, ineffective, or even harmful, for others. Even if an experimental treatment is at least mildly effective for an entire population, the standard of care may still be preferred for some individuals if, for example, the experimental treatment is very expensive and there is little difference in effectiveness between the two [Song and Pepe, 2004]. Thus, it is desirable to know which individuals in a population, if any, will respond well to a particular treatment. In particular, the identification of the characteristics which lead to these individuals showing an enhanced response is of interest, as this may allow future members of the population to be assigned the treatment which will benefit them the most.

In recent years, many authors have proposed methods which use randomized clinical trial or observational data to obtain a set of "rules" based on patient information, which can subsequently be used to help ensure that future individuals in the population are assigned the treatment which is best for them [Friedman and Fisher, 1999, Negassa et al., 2005, Gunter et al., 2007, Su et al., 2008, 2009, Brinkley et al., 2010, Cai et al., 2011, Foster et al., 2011, Janes et al., 2011, Lipkovich et al., 2011, Qian and Murphy, 2011, Zhao et al., 2011, Imai and Ratkovic, 2012, Zhang et al., 2012]. The resulting "rules" may vary widely in form and complexity, from simple regions of the design space, such as $x_3 > 5$ or $\{x_1 > 0.5, x_6 < 1\}$, to more complex inequalities, such as $f(\boldsymbol{x}) > c$, where c is some constant and f is some nontrivial p-to-1 function of the covariate vector $\boldsymbol{x}$.

In this paper, we limit our discussion to cases where the outcome is continuous, and only two treatment options are available. We are interested in cases where p, the number of baseline covariates, is moderate, e.g. 5 to 100. We consider the use of randomized clinical trial (RCT) data to select a simple treatment "regime" [Zhang et al., 2012] which, if followed by the entire population, leads to the best expected outcome. Potential regimes are restricted to those which involve assigning one treatment to individuals who are in a region, say A, of the covariate space, and assigning the other treatment to those individuals in A^c. It is desirable that these regions be simple, and depend on only a limited number of covariates, so potential regions are limited to contiguous subsets of the covariate space defined by one, two and three variables, such as $\{x_1 > 0, x_2 > 0\}$ or $\{x_3 < 5, x_4 > 0, x_7 < 1\}$. These simple regions are easy to understand, and allow for future treatment decisions to potentially be faster, less expensive or less invasive, as they require only a limited amount of necessary patient information. In addition to providing a "nice" functional form, limiting the number of covariates that define the regions allows one to potentially identify the covariates which most strongly affect how a patient will respond to treatment.

The remainder of this chapter is as follows. In Section 2 we describe the proposed method and outline an algorithm for implementation. In Section 3 we present the results of a simulation study, and in Section 4, we discuss the application of the proposed methods to a prehypertension RCT data set.

3.2 Identifying subgroups using the average value function

Suppose we have independent observations $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ from the general model

$$y_i = h(\boldsymbol{x}_i) + (T_i - \pi)g(\boldsymbol{x}_i) + \epsilon_i, \tag{3.1}$$

where y is a continuous outcome, g and h are unknown functions, T is a binary treatment indicator, π is the treatment randomization probability, and $\epsilon_1, \ldots, \epsilon_n$ are iid errors with mean zero and variance σ^2. Without loss of generality, assume that higher values of y represent an improved response. We wish to estimate a subregion, $\hat{A}$, of the covariate space with which to define our treatment regime. In future populations, only individuals in region $\hat{A}$ would receive treatment ($T = 1$), with those in $\hat{A}^c$ receiving the standard of care ($T = 0$). We wish to select the region $\hat{A}$ which maximizes the expected response under this regime. This expectation is sometimes referred to as the average Value [Sutton and Barto, 1998, Gunter et al., 2007]. Note that $g(\boldsymbol{x}_i)$ is the treatment effect for subject i, so if we had no restrictions on $\hat{A}$ and g were known, the best regime would be to treat all individuals with $g(\boldsymbol{x}_i) \geq 0$. The functions $g(\boldsymbol{x})$ and $h(\boldsymbol{x})$ may be complex, potentially involving non-linearities and interactions, so the approach we take is to use non-parametric methods to estimate g and h. These estimates are then used to select the "optimal" region $\hat{A}$, with the restriction that $\hat{A}$ has to be simple.

3.2.1 Nonparametric estimation of g and h

The following iterative approach is used to estimated the unknown functions h and g in (3.1):

(i) Fit the model $y = h(\boldsymbol{x})$ to obtain the initial estimate of h, $\hat{h}^{(1)}$.

(ii) Fit the model $\frac{1}{T-\pi}(y - \hat{h}^{(k)}(\boldsymbol{x})) = g(\boldsymbol{x})$ to obtain the k^{th} estimate of g, $\hat{g}^{(k)}$, $k \geq 1$.

(iii) Fit the model $y - (T-\pi)\hat{g}^{(k)}(\boldsymbol{x}) = h(x)$ to obtain the $(k+1)^{th}$ estimate of h, $\hat{h}^{(k+1)}$, where $k \geq 1$.

(iv) Iterate between steps (ii) and (iii) until $\sum_{i=1}^{n}\left[y_i - \hat{h}^{(k)}(\boldsymbol{x}_i) - (T-\pi)\hat{g}^{(k)}(\boldsymbol{x}_i)\right]^2$ changes by less than a prespecified small number.

There are many possible choices of model or algorithm for estimating g and h in steps (ii) and (iii), such as multivariate adaptive regression spline (MARS) [Friedman, 1991] and Random Forests [Breiman, 2001]. One may wish to choose the “convergence threshold” in step (iv) above differently depending on which estimation method is chosen. For instance, in our experience, a threshold of around 10^{-5} can generally be achieved within only a few iterations for methods such as MARS and generalized additive models. For Random Forests, the amount by which the sum of squares in step (iv) changes remains somewhat constant, regardless of how many iterations have been performed, most likely because of the random nature of this method. Thus, in this case, we instead continue until 60 iterations have been performed, as in our experience this is more than enough to obtain good estimates of $g(\boldsymbol{x})$ and $h(\boldsymbol{x})$.

3.2.2 Selecting a subgroup for fixed g and h

Using notation similar to Zhang et al. [2012], let y_{1i} and y_{0i} be the potential responses given that subject i received treatment or the standard of care respectively, so that $y_i = y_{1i}T_i + y_{0i}(1 - T_i)$. Let $y_i^*(A) = y_{1i}I(\boldsymbol{x}_i \in A) + y_{0i}(1 - I(\boldsymbol{x}_i \in A))$ be the potential outcome for a future subject under this "treat-if-in-A" regime for any region A. After some simple algebra, we have

$$\begin{aligned} E\Big[y_i^*(A)\Big] &= E\Big[E\Big(y_i^*(A)\Big|\boldsymbol{x}_i, A\Big)\Big] \\ &= E\Big[h(\boldsymbol{x}_i)\Big] - E\Big[\pi g(\boldsymbol{x}_i)\Big] + E\Big[2\pi g(\boldsymbol{x}_i)I(\boldsymbol{x}_i \in A)\Big]. \end{aligned} \tag{3.2}$$

Note that only the last term in (3.2) involves A, so maximizing the expected value of $y_i^*(A)$ with respect to A amounts to maximizing

$$E_{\boldsymbol{X}}\Big[g(\boldsymbol{x}_i)I(\boldsymbol{x}_i \in A)\Big]. \tag{3.3}$$

Thus, given function g, one estimator for (3.3) is the sample average:

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i)I(\boldsymbol{x}_i \in A),$$

which, after multiplying by n, can be rewritten as

$$\sum_{i:\boldsymbol{x}_i \in A} g(\boldsymbol{x}_i). \tag{3.4}$$

The chosen subgroup, $\hat{A}$, is that which maximizes (3.4). In practice, one may wish to consider the inclusion of a nonzero offset in (3.4), as in our experience this can help to better identify truly positive responders. Specifically, one could replace (3.4) with

$$\sum_{i:\boldsymbol{x}_i \in A} [g(\boldsymbol{x}_i) - \delta], \tag{3.5}$$

where $\delta \neq 0$. Selection of the offset δ is considered below.

We consider one, two and three-dimensional regions of the general form $\left\{x_j \gtrless c_j\right\}$,or $\left\{x_j \gtrless c_j\right\} \cap \left\{x_k \gtrless c_k\right\}$ or $\left\{x_j \gtrless c_j\right\} \cap \left\{x_k \gtrless c_k\right\} \cap \left\{x_l \gtrless c_l\right\}$ as candidates for the region $\hat{A}$, where $\gtrless$ indicates either $\geq$ or $<$ and j, k and l are distinct. In addition, we consider the complements of these regions.

Once the final converged estimates, $\hat{h}$ and $\hat{g}$, are obtained, the optimal region $\hat{A}$ can be found by replacing g with $\hat{g}$ in (3.4) or (3.5) and maximizing with respect to A. For the remainder of the paper, this subgroup identification method will be referred to as the *Average Value* (AV) approach.

Note that for just the three-variable candidate regions specified above, there are $\binom{p}{3}$ unique combinations of covariates, $2^3 = 8$ unique ways to assign directions $\geq / <$ to x_j, x_k and x_l, and as many as $n-1$ unique cutpoints for each covariate (if observed values are unique). Thus, the Average Value procedure will often involve the evaluation of a very large number of regions, making it very computationally expensive. Therefore, a modified version of the procecure is employed in our simulations and example data analysis. Specifically, to decrease the number of candidate groups, we consider a rough, evenly-spaced grid of 10 to 20 cutpoints, rather than all observed values as candidates for the cutpoint c_j for covariate j, $j = 1, \ldots, p$. Additionally, instead of considering all one, two and three-dimensional regions simultaneously, we employ a "stepwise" approach. This approach is as follows:

1. Evaluate all candidate one-dimensional regions, and select the best M_{1D} re-

gions. Let B_{1D} be the set of unique covariates which define these "best" one-dimensional regions.

2. Evaluate all candidate two-dimensional regions in which one of the dimensions is defined by a member of B_{1D} and the other is any other covariate (that may or may not be in B_{1D}), and select the best M_{2D} regions. Let B_{2D} be the set of unique pairs of covariates which define these "best" two-dimensional regions. Note that B_{1D} is only used to define *which* covariates are allowed to define one of the dimensions of the two-dimensional regions. All possible directions (i.e. $<$ or $\geq$) and cutpoints are considered for these covariates when evaluating two-dimensional regions.

3. Evaluate all candidate three-dimensional regions in which two of the dimensions are defined by a pair contained within B_{2D} and the other is any other covariate (that may or may not be in B_{2D}), and select the best three-dimensional region. As in step 2, B_{2D} is only used to define which covariates are allowed to define two of the dimensions of the three-dimensional groups, so all possible directions (i.e. $<$ or $\geq$) and cutpoints are considered for these covariates when evaluating three-dimensional regions.

4. Identify $\hat{A}$, the best overall region.

It should be noted that, if one wishes to restrict the minimum or maximum size of candidate subgroups, this can be done by simply removing groups outside of this size range from consideration. This may also help to improve the computational efficiency of the AV method, as it further decreases the number of subgroups which must be evaluated.

3.2.3 Evaluation of the region $\hat{A}$

The proposed method always identifies a region, so it is important to have a method by which to evaluate the strength of the selected region. For this purpose, we consider the metric proposed by Foster et al. [2011]:

$$Q(A) = E(y|T = 1, \boldsymbol{x} \in A) - E(y|T = 0, \boldsymbol{x} \in A) - [E(y|T = 1) - E(y|T = 0)], \quad (3.6)$$

which is a measure of the enhanced treatment effect in A relative to the average treatment effect. We consider six methods for estimating (3.6), which we briefly describe below. Additional details are given by Foster et al. [2011].

Method 1. Resubstitution. Replace the four conditional expectations in (3.6) with the observed means in the data and use these obtain an estimate of $Q(A)$:

$$\begin{aligned}\hat{Q}(A)_{RS} = {} & \frac{\sum_{i=1}^{n} y_i I(\boldsymbol{x}_i \in A, T_i = 1)}{\sum_{i=1}^{n} I(\boldsymbol{x}_i \in A, T_i = 1)} - \frac{\sum_{i=1}^{n} y_i I(\boldsymbol{x}_i \in A, T_i = 0)}{\sum_{i=1}^{n} I(\boldsymbol{x}_i \in A, T_i = 0)} \\ & - \left[\frac{\sum_{i=1}^{n} y_i I(T_i = 1)}{\sum_{i=1}^{n} I(T_i = 1)} - \frac{\sum_{i=1}^{n} y_i I(T_i = 0)}{\sum_{i=1}^{n} I(T_i = 0)} \right]. \end{aligned} \quad (3.7)$$

As noted by Foster et al. [2011], the Resubstitution method re-uses the data which were used to estimate $\hat{A}$. Thus, this estimate is expected to be positively biased.

Method 2. Simulate new data. The goal of this method is to obtain new data which “look like” the original data, but are independent of the original data, thereby reducing the bias of the resulting estimate. This process could be repeated many times, and each time (3.7) could be recalculated using the new data. The simulate new data estimate could be found by averaging these resubstitution estimates. As in Foster et al. [2011], we avoid actually simulating new data by instead replacing y_i by $\hat{y}_i = \hat{h}(\boldsymbol{x}_i) + (T_i - \pi)\hat{g}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$ in (3.7). This estimate is denoted $\hat{Q}(A)_{SND}$. This method also tends to be positively biased, but is generally less biased than the RS approach. It should be noted that, though we do not actually simulate new data in this method, we use the name "simulate new data" in order to be consistent with Foster et al. [2011].

Method 3. Mean $\hat{g}$. For a given sample, under model (3.1), the empirical version of (3.6) is equivalent to $\frac{1}{|A|}\sum_{i:\boldsymbol{x}_i \in A} g(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i)$, where $|A|$ is the number of individuals in the region A. Thus, $\hat{g}$ can be used to obtain an estimate of (3.1):

$$\hat{Q}(A)_{\hat{g}} = \frac{1}{|A|}\sum_{i:\boldsymbol{x}_i \in A} \hat{g}(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{g}(\boldsymbol{x}_i). \qquad (3.8)$$

This method is similar to the SND method, and will generally have a similar amount of bias. It should be noted that for "paired" data, in which each treated observation has a corresponding identical (with respect to covariate values) control observation, this estimate will be exactly equal to the SND estimate.

Methods 4-6. Bootstrap bias correction. We also consider the bootstrap bias correction discussed in [Foster et al., 2011]. For a given region A, the bias of an estimate, $\hat{Q}(A)$, of (3.6) is $\hat{Q}(A) - Q(A)$. Following the arguments provided in [Foster et al., 2011], it is possible to obtain an approximation to this bias using bootstrap data sets. Let sample l represent a bootstrap sample taken with replacement from the observed data. Using this sample, the Average Value approach is implemented, providing new estimates $\hat{h}^{(l)}$, $\hat{g}^{(l)}$ and a new region

$A^{(l)}$. A new estimate $\hat{Q}^{(l)}(A^{(l)})$ of (3.6) is then obtained from one of the above methods, and can be viewed as an approximate estimate of $\hat{Q}(A)$. Additionally, the observed data and corresponding function estimates $\hat{h}$ and $\hat{g}$ may be used along with the new region $A^{(l)}$ to calculate $\hat{Q}(A^{(l)})$, which can be viewed as an approximate estimate of $Q(A)$. Thus, a bootstrap approximation of the bias is $\hat{Q}^{(l)}(A^{(l)}) - \hat{Q}(A^{(l)})$. This process is repeated many times (say, L), and the bias of $\hat{Q}(A)$ is approximated by $\frac{1}{L}\sum_{l=1}^{L}(\hat{Q}^{(l)}(A^{(l)}) - \hat{Q}(A^{(l)}))$. The resulting estimate may be used to adjust estimates from any of the above three methods, i.e. the adjusted $\hat{Q}(A) = \hat{Q}(A) - \frac{1}{L}\sum_{l=1}^{L}(\hat{Q}^{(l)}(A^{(l)}) - \hat{Q}(A^{(l)}))$. These adjusted RS, SND and $\hat{g}$-based estimates are called Methods 4, 5 and 6 respectively.

3.2.4 Selection of δ

If one wishes to consider an offset, δ, a number of options exist. In this paper, we use the offset to improve the performace of our method, and we select the smallest value δ such that $(-\delta, \delta)$ contains at least 50% of the estimated treatment effects, $\hat{g}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, thus forcing the estimated subgroup to contain fewer than half of the observations. This value tends to be large, and was selected because we wish to aggressively search for enhanced individuals. It should be noted that, though we use a positive, nonzero δ, we consider anyone with a treatment effect greater than zero to be enhanced. If one instead considers “enhanced” individuals to be those beyond some known, nonzero, minimal meaningful treatment effect (say, c), one could instead define the offset to be $\delta + c$, where δ is still nonzero. Alternatively, if one wishes for the selected subset to be of a specific size, the value of δ could be chosen accordingly. If one wishes to be less aggressive, an offset need not be used.

3.3 Simulations

To evaluate the performance of the proposed method a simulation study was undertaken. In this study, the proposed method is compared to the Virtual Twins [Foster et al., 2011] approach, which is another two-stage subgroup identification procedure. In the first stage of the Virtual Twins procedure, y is used as the outcome in a Random Forest, with the covariates and treatment indicator being used as predictors. This Random Forest is used to obtain estimates of the two potential outcomes, y_{1i} and y_{0i}, for each subject, and the estimated treatment effect $\hat{y}_{1i} - \hat{y}_{0i}$ is calculated for each subject. In the second stage, the estimated treatment effects are used as the outcome in a single regression tree, and the identified subgroup consists of all terminal nodes for which the estimated treatment effect (from the single regression tree, not the Random Forest) is beyond some pre-defined "enhancement" threshold.

In this simulation study, we consider six different cases, as shown below:

1. $g(\boldsymbol{x}) = \frac{5}{\sqrt{2}}(x_1 + x_2)$

2. $g(x) = \begin{cases} 0.5 + 10\min(|x_1|, |x_2|) & \text{if } x_1 > 0.5 \text{ and } x_2 > 0.5 \\ -0.5 & \text{if } x_1 < -0.5 \text{ or } x_2 < -0.5 \\ \min(|x_1 + 0.5|, |x_2 + 0.5|) - 0.5 & \text{otherwise} \end{cases}$

3. $g(\boldsymbol{x}) = 20I(x_1 > 0, x_2 > 0)\min(|x_1|, |x_2|)$

4. $g(\boldsymbol{x}) = 35I(x_1 > 0, x_2 > 0)$

5. $g(\boldsymbol{x}) = 5$

6. $g(\boldsymbol{x}) = 0.$

In all cases, at most 2 of the p variables determine $g(\boldsymbol{x})$. In case 1, treatment effects are generated by summing values of the first two covariates. Because the covariates are normally distributed with mean zero, the resulting treatment effects have a distribution which is symmetric about zero. Thus, there is no clearly separated "enhanced" group of individuals who are different from the rest of the population. However, individuals with $x_1 + x_2 > 0$ show a positive expected response to treatment and those with $x_1 + x_2 < 0$ show a negative expected response to treatment. Cases 2-4 have clearly defined enhanced individuals present. In case 2, there is a group of "nonresponders" whose values for g vary slightly around zero, and a group of "responders", whose values for g vary around some nonzero mean treatment effect. Case 3 is similar to case 2, but nonresponders show no effect of treatment whatsoever, rather than small effects centered at zero. In case 4, nonresponders again have a constant zero treatment effect, and responders have a constant nonzero treatment effect. Cases 5 and 6 are two variants on a null case. In case 5, the treatment effect is a non-zero constant for all individuals, so no "enhanced" region exists, or alternatively, everyone is "enhanced." In case 6, the treatment effect is exactly zero for all members of the population. Specific data generation schemes for all six cases are given below. For each case, 100 data sets of size $n = 500$ were generated from the model:

$$y_i = 30 + 5x_{1i} + 5x_{2i} - 5x_{7i} + T_i g(\boldsymbol{x}_i) + \epsilon_i,$$

where x's are iid standard normal, ϵ's are iid normal with mean zero and variance 100 and are independent of the x's. In all cases, we consider a total of 10 variables (8 of which may be considered "noise" variables) in our analysis. For cases 1-4, three-dimensional plots of x_1, x_2 and g are given in Figure 3.1 and histograms of g from all

100 data sets can be found in Figure 3.2. The "true" enhanced region in each case consists of all individuals for whom $g(\boldsymbol{x}_i) > 0$. Thus, in case 1 we expect $\frac{1}{2}$ of the subjects to be enhanced, in cases 2-4 we expect $\frac{1}{4}$ of the population to be enhanced, in case 5 the true region is all individuals, and in case 6 the true region is empty.

For the Average Value approach, only subgroups of size 20 or larger were considered. It should be noted that this value is somewhat arbitrary. Additionally, for the stepwise subgroup search, the top 10 ($M_{1D} = 10$) of the one-dimensional regions were used to identify covariates for B_{1D} and the top five of the two-dimensional groups of the form $\left\{x_i \underset{<}{\geq} c_i, x_j \underset{<}{\geq} c_j\right\}$ (and top five groups of the form $\left\{x_i \underset{<}{\geq} c_i, x_j \underset{<}{\geq} c_j\right\}^c$) were used to identify covariate pairs for B_{2D} (when combined, gives $M_{2D} = 10$). Candidate cutpoints for each covariate were the corresponding $0, 5, 7.5, \ldots, 95$ percentiles for the one-dimensional search, $0, 5, 10, \ldots, 95$ percentiles for the two-dimensional search and $0, 5, 20, 35, 50, 65, 80, 95$ percentiles for the three-dimensional search. It should be noted that the 0^{th} quantiles were included as candidate cutpoints to allow for the identification of subgroups of less than three variables, since $\geq 0^{th}$ percentile means the corresponding variable is useless. For both the Average Value and Virtual Twins procedures, 20 bootstrap data sets were used to obtain the bias-corrected estimates. As mentioned above, for the Average Value procedure, we selected an offset δ such that $(-\delta, \delta)$ contains at least 50% of the estimated treatment effects. This offset value was also used as the "enhancement" threshold for the Virtual Twins procedure. For the Average Value procedure, the unknown functions g and h were estimated using a simple average MARS estimates and Random Forest estimates, as this approach was found to perform better than either method alone in our simulations. These estimates were obtained using the R functions *randomForest* and *mars* with default settings.

For each case, to assess the ability of the methods to identify the true underlying subgroup, we calculate the average number of individuals with a true positive treatment effect, the average size of the identified region, the average sensitivity, specificity, positive and negative predictive values for the identified regions, the proportion of the time in which the correct covariates are included in the identified regions, and the proportion of the times the identified subgroup is defined using only the correct covariates. In addition, for each case we calculate the average values of $Q(A)$, $Q(\hat{A})$ and all the estimates of $Q(\hat{A})$ discussed in Section 2.2.3. In the calculation of sensitivity, specificity, positive predictive value and negative predictive value, we consider the "true" region to be all individuals for whom $g(\boldsymbol{x}_i) > 0$ and the "estimated" region to be all individuals in $\hat{A}$.

Table 3.1, shows that the Average Value approach tends to identify larger subgroups than Virtual Twins, especially when there exists a subgroup of patients with especially large treatment effects, such as in Cases 3 and 4. Because of this, the Average Value approach tends to have slightly better sensitivity and slightly lower positive predictive value than Virtual Twins. This tendency to select larger groups also leads to slightly lower specificity for the Average Value approach compared to Virtual Twins, especially when some very enhanced individuals exist (Cases 3 and 4), as the Average Value approach tends to identify too many individuals as being enhanced in these cases. Both methods tend to identify too few individuals as being enhanced in situations where there are many individuals with small to moderate treatment effects, such as in Cases 1, 2 and 5; however, this is to be expected, as a large offset was selected in order to aggressively search for very enhanced individuals. The Average Value method is generally very successful at identifying regions which depend on the true important covariates, and this success appears to be less sensitive to changes in scenario than that of the Virtual Twins approach. The Virtual Twins approach more frequently identifies regions which depend only on the correct covariates. Both of these trends are most likely a result of the Average Value method's tendency to select three-dimensional regions, regardless of the true underlying dimension.

From Table 3.2, we can see that the Virtual Twins procedure tends to identify more enhanced regions than the Average Value procedure. This is most likely due to the fact that Virtual Twins tends to identify fewer subjects as being enhanced than the Average Value procedure. The uncorrected estimates of $Q(\hat{A})$ tend to be less biased for Virtual Twins than for the Average Value procedure, most likely because the Average Value procedure selects less enhanced regions, and as noted by Foster et al. [2011], these estimates tend to be more biased when $Q(\hat{A})$ is small (or zero). As expected, the uncorrected SND estimates are less biased than the RS estimates for both procedures. Also, for the Average Value approach, the Mean $\hat{g}$ estimate appears to be a slight improvement over the SND estimate. The bias correction appears to work better for the Average Value method, showing less of a tendency to overcorrect than with Virtual Twins, perhaps because the estimates of $Q(\hat{A})$ tend to be more biased for the Average Value procedure than for Virtual Twins.

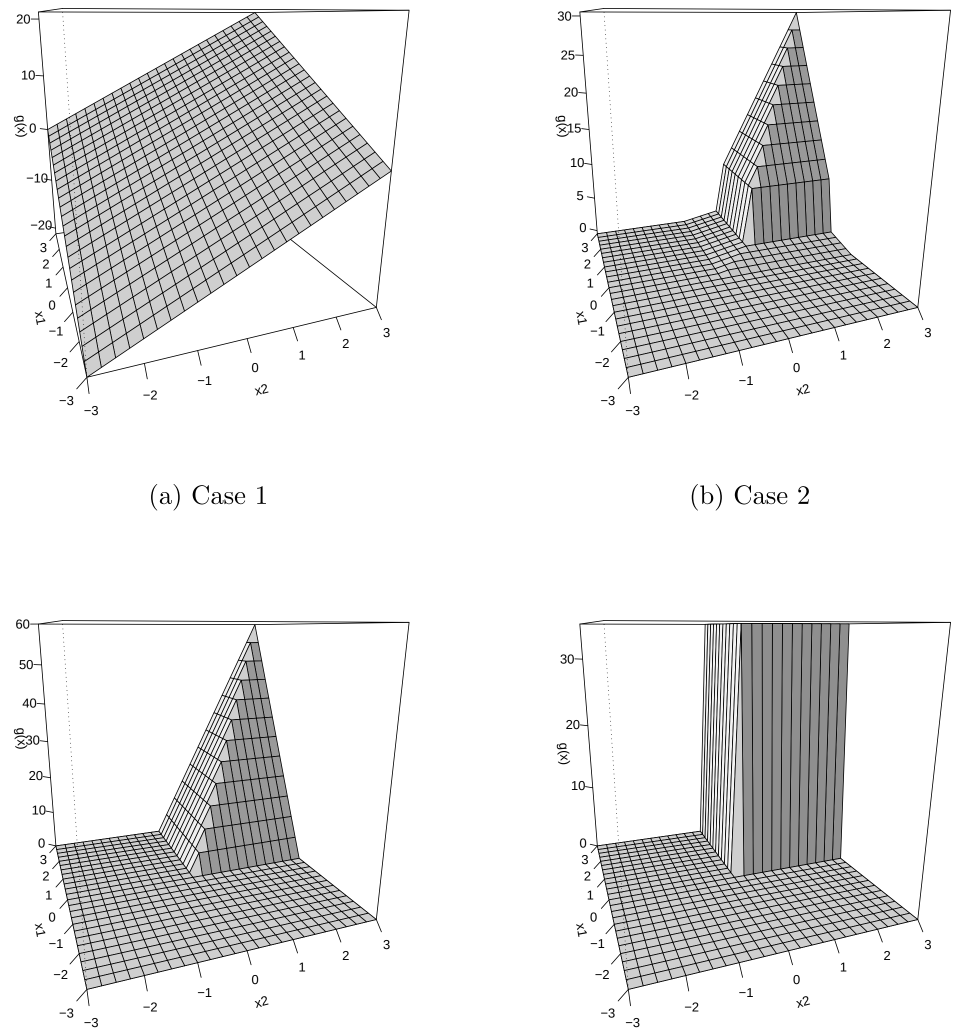
It is difficult to identify one estimate as the best performer for all cases for either method. For non-null cases (1-4), the uncorrected SND estimates appear to be best for Virtual Twins, while the bias-corrected RS estimates generally perform best for the Average Value method. Though it tends to overcorrect, the bias-corrected Mean $\hat{g}$ estimate also seems promising, especially for null cases 5 and 6. For the Virtual Twins procedure, the bias-corrected RS estimates appear to be best for cases 5 and 6.

In our experience, the Virtual Twins procedure has a tendency to identify subgroups which consist of two or more disjoint regions, whereas the Average Value method is designed to identify only contiguous regions. Because of this, subgroups identified by the Average Value method will generally be simpler and easier to use than those identified by Virtual Twins. Moreover, the results of this simulation study suggest that further restricting the form of potential subgroups has only a very mild negative impact on performace (compared to Virtual Twins). Thus, we believe the Average Value procedure is a very viable alternative to Virtual Twins.

Table 3.1: Simulation study results: subgroup identification performance

	True #						Incl.	Only
Scenario	Responders[1]	Size	Sens.	Spec.	PPV	NPV	x_1, x_2	x_1, x_2
Case 1								
AV:	249.94	98.84	0.36	0.97	0.93	0.61	0.99	0.11
VT:	249.94	95.12	0.36	0.98	0.92	0.61	0.94	0.43
Case 2								
AV:	125.27	103.08	0.33	0.84	0.42	0.79	1.00	0.10
VT:	125.27	78.88	0.37	0.91	0.62	0.82	0.81	0.12
Case 3								
AV:	125.27	171.37	0.69	0.77	0.53	0.89	1.00	0.21
VT:	125.27	118.79	0.62	0.89	0.73	0.88	1.00	0.43
Case 4								
AV:	125.27	204.32	0.99	0.79	0.63	1.00	1.00	0.53
VT:	125.27	129.93	0.97	0.98	0.96	0.99	1.00	1.00
Case 5								
AV:	500	239.30	0.48	-	1.00	-	-	-
VT:	500	231.64	0.46	-	1.00	-	-	-
Case 6								
AV:	0	76.74	-	0.85	-	1.00	-	-
VT:	0	49.42	-	0.90	-	1.00	-	-

[1] True responders defines as those with $g(\boldsymbol{x}_i) > 0$.
AV and VT indicate Average Value and Virtual Twins respectively.
Sample size for each case is 500.



(a) Case 1

(b) Case 2

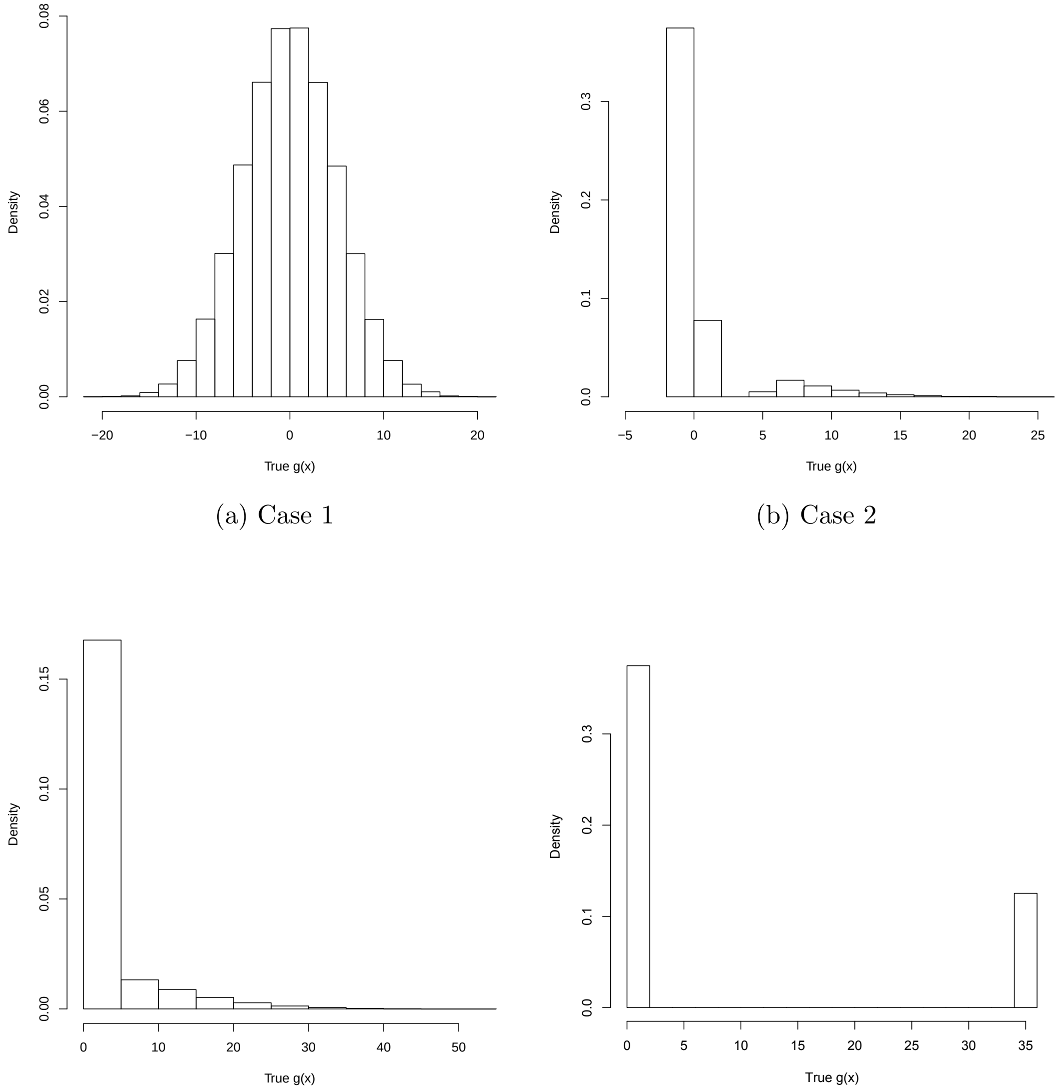
(c) Case 3

(d) Case 4

Figure 3.1: Plot of True $g(\boldsymbol{x})$ for Cases 1-4

3.4 Application to randomized clinical trial data

The proposed methods were applied to data from the Trial of Preventing Hypertension (TROPHY) [Julius et al., 2006]. This study included participants with prehypertension, meaning that all participants had either an average systolic blood pressure of 130 to 139 mm Hg and diastolic blood pressure of no more than 89mm Hg



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 3.2: Histogram of True $g(\boldsymbol{x})$ for Cases 1-4

for the three run-in visits (before randomization), or systolic pressure of 139 mm Hg or lower and diastolic pressure between 85 and 89 mm Hg for the three run-in visits. These subjects were randomly assigned to receive either two years of candesartan (a hypertension treatment) or placebo, followed by two years of placebo for all subjects. Subjects had return visits at 1 and 3 months post-randomization, and every 3 months

Table 3.2: Simulation study results: $Q(\hat{A})$ estimation performance

			$\hat{Q}(\hat{A})$			Bias-Corrected $\hat{Q}(\hat{A})$		
Scenario	$Q(A)$	$Q(\hat{A})$	RS	SND	Mean $\hat{g}$	RS	SND	Mean $\hat{g}$
Case 1								
AV:	4.00	5.43	8.00	6.63	6.63	4.97	2.79	3.51
VT:	4.00	6.19	7.92	6.29	-	4.00	1.16	-
Case 2								
AV:	3.03	1.38	6.10	4.38	4.11	2.82	0.19	0.65
VT:	3.03	4.31	6.45	4.91	-	1.62	-1.16	-
Case 3								
AV:	7.02	3.63	5.63	4.53	4.37	3.29	1.53	1.92
VT:	7.02	7.71	9.01	7.17	-	5.51	2.67	-
Case 4								
AV:	26.23	13.31	14.36	11.72	11.28	13.23	9.98	10.32
VT:	26.23	24.95	25.01	23.20	-	24.28	22.20	-
Case 5								
AV:	0.00	0.00	2.69	1.87	1.79	0.54	-0.90	-0.61
VT:	0.00	0.38	2.49	2.20	-	-1.47	-2.56	-
Case 6								
AV:	0.00	0.00	6.12	4.33	3.85	2.33	-0.43	-0.09
VT:	0.00	0.41	2.75	2.29	-	-0.77	-2.05	-

AV and VT indicate Average Value and Virtual Twins respectively.
Sample size for each case is 500.

thereafter until month 24. In year 3, clinic visits were at 25 and 27 months, and then every third month thereafter until the end of the study. The study produced analyzable data on 772 subject, with 391 coming from the candesartan group and 381 from the placebo group. Baseline measurements included age, gender, race (white, black or other), weight, body-mass index (BMI), systolic and diastolic blood pressures, total cholesterol, high density lipoprotein cholesterol (HDL), low density lipoprotein cholesterol (LDL), HDL:LDL ratio, triglycerides, fasting glucose, total insulin, insulin:glucose ratio and creatinine, with the insulin:glucose ratio being dropped from our analysis due to extremely high correlation (≈ 0.98) with total insulin. The endpoint of interest in the original study was the binary variable development of stage 1 hypertension (see Julius et al. [2006] for details); however, for the purpose of this paper we consider the continuous variable systolic blood pressure as the outcome variable. Specifically, the outcome in our analysis is blood pressure (systolic) at 12 months post-randomization.

It should be noted that at 12 months post-randomization there was some (approximately 20%) missing data in the outcome due to patient dropout and patients developing hypertension (the endpoint in the original study). For our analysis, because the endpoint (hypertension) was defined based only on *observed* blood pressure measurements, missing data due to patients experiencing the event were assumed to be missing at random. There was also a small amount of missingness in the baseline covariates, with the largest fraction of missing for any covariate being 4.3%. All missing values were imputed using SAS PROC MI (SAS Institute Inc., Cary, NC). The imputation model included all baseline covariates and all blood pressure measurements up to 12 months post-randomization, stratified by treatment and gender. Because the proposed methods have not yet been extended to data with missing values, only a single imputation was perfomed.

There are three very large and influential outliers in the covariate values. Thus, Random Forests, rather than an average of Random Forests and MARS, was used to estimate the unknown functions g and h, as it is less sensitive to outliers than MARS. Additionally, insulin, glucose, HDL, LDL, HDL:LDL ratio and triglycerides were noticeably skewed, so these covariates were log-transformed for the analysis.

In our analysis, we used the following settings for the Average Value method: (1) the percentiles of covariates used as cutpoints in the three-dimensional search were changed to 0, 7.5, 15, $\ldots$, 90, (2) all two-dimensional regions were considered in the stepwise search procedure (so that B_{1D} contained all the baseline covariates), (3) the top 50 two-dimensional groups (and the top 50 complement groups) were used to identify covariate pairs for B_{2D} (so $M_{2D} = 100$), and (4) the Random Forest included 2000 trees. All other settings were the same as those in the simulations.

A histogram of the estimated treatment effects is given in Figure 3.3. The very high percentage of positive predicted treatment effects suggests that candesartan is widely effective for treating prehypertension. Thus, in this case, it may be more interesting to identify the small subgroup of individuals who shouldn't receive treatment. As a result, no offset was used in this analysis ($\delta = 0$), and $\hat{A}$ was redefined as the region which minimizes (3.5).

The identified region is $\hat{A} = \{\text{HDL:LDL ratio} < 0.38,\ \text{HDL cholesterol} < 46.02,$ $\text{total insulin} \geq 25.11\}$, and contains 20 subjects. Thus, the Average Value method suggests a treatment regime where individuals in this region receive placebo and all others receive candesartan. Estimates of $Q(\hat{A})$ were -1.63, -8.14 and -9.75 for the RS, SND and Mean $\hat{g}$ methods respectively, with bias corrected values being 3.92, 0.35 and -4.24. The bias corrected estimates are closer to zero than the uncorrected estimates, and are fairly small in magnitude, suggesting that individuals in the identified region may have essentially no response to treatment, rather than a large negative response.

It should be noted that, due to the random nature of Random Forests, results may vary slightly depending on which seed is chosen for the estimation of the functions g and h. The above analysis was repeated using a different random seed, and a slightly different subgroup was identified; however, the subgroup was again defined using insulin and two of the cholesterol measures, and contained some, but not all of the same individuals. The above analysis was also performed without the three large outlying observations in the covariates, and again a region based on cholesterol

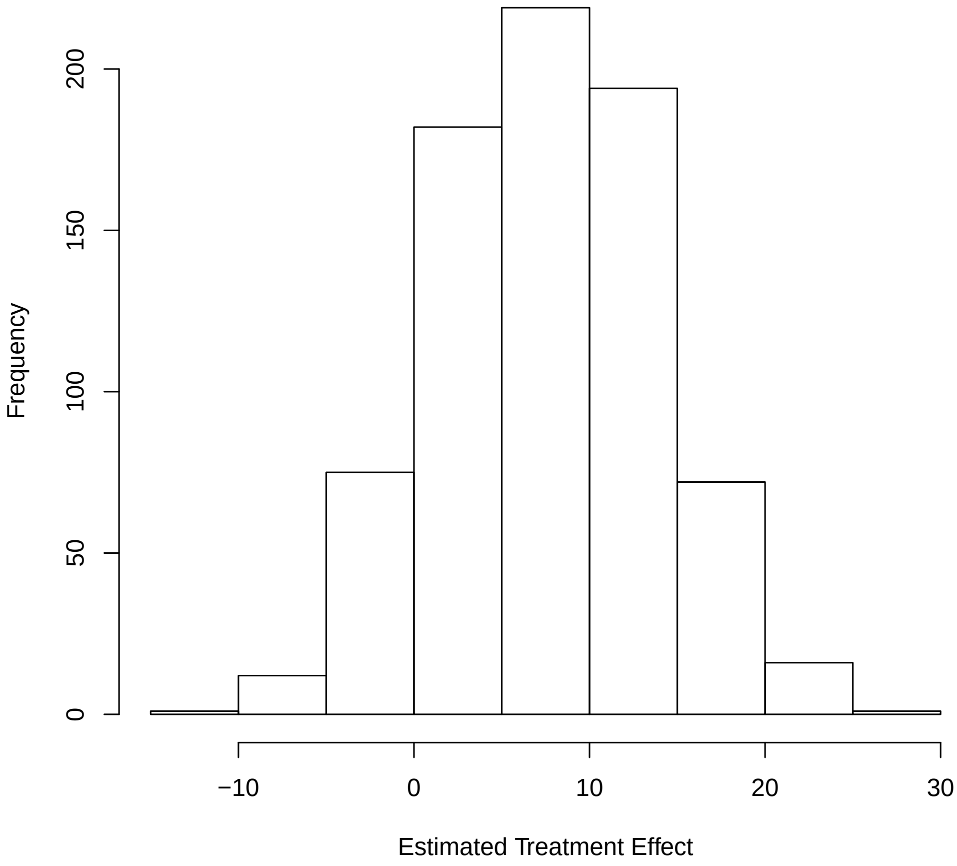


Figure 3.3: Histogram of $\hat{g}(\boldsymbol{x})$ for TROPHY Data

measures and insulin was identified. Given the relatively small magnitude of the bias-corrected estimates of $Q(\hat{A})$, it is also possible that the selection of different covariates and subjects with different seeds is due to the fact that no "true" subgroup exists.

3.5 Discussion

We proposed a method which uses randomized clinical trial data to identify simple sets of "rules" based on patient information, which can be used to define future treatment regimes. The method was found to be very effective at identifying truly important covariates in simulations, but had a tendency to identify larger, and therefore less enhanced subgroups than Virtual Twins [Foster et al., 2011]. Though slightly less enhanced, the subgroups identified by the Average Value procedure were generally comparable to those from Virtual Twins, and have the added advantage of being simpler, and therefore easier to interpret, which may lead to them seeing more real-world use.

Due to slight differences between the proposed method and Virtual Twins, it is difficult to know if the results in the simulation section are directly comparable. For instance, an offset of, say 10 for the Average Value procedure may not be equivalent to an "enhancement" threshold of 10 in Virtual Twins. Additionally, a direct comparision may not be fair because, as previously mentioned, the Average Value procedure does not allow the same subgroup complexity as Virtual Twins.

Though very effective at identifying truly important covariates, the proposed method tends to select three-dimensional regions, even when the true underlying region is of fewer dimensions. Thus, it may be interesting to consider some form of pruning, as is done for classification and regression trees. This would be particularly interesting if one wished to consider subgroups of more than three dimensions. Alternatively, it may be interesting to consider incorporating a penalty based on the number of covariates to the objective function. It is possible that the inclusion of such a penalty could help the Average Value method to more frequently identify regions of the correct dimension.

The general strategy used in this paper is to first estimate $g(\boldsymbol{x}_i)$, and then find the region A which maximizes $\sum_{i \in A} \hat{g}(\boldsymbol{x}_i)$, where we restrict $\hat{A}$ to have a certain simple form. Other simple forms of $\hat{A}$ could also be considered. For example, $\hat{A}$ could be defined by $\{\boldsymbol{x} : \boldsymbol{\beta}^T \boldsymbol{x} > c\}$, with further simplification possible if the number of non-zero β_j's was also restricted.

It should be noted that the chosen value of the offset δ can strongly impact the size of the estimated subgroup. Thus, it may be possible to improve the performance of the Average Value approach by using alternative data-adaptive methods for selecting δ.

Because of the computationally expensive nature of the Average Value method, we considered a more computationally efficient "stepwise" version of the procedure in our data analysis and simulations. It may be of interest to consider less greedy methods for increasing the speed of the procedure.

CHAPTER 4

Permutation testing for treatment-covariate interactions

4.1 Introduction

In clinical trials, a common goal is to search for subgroups of enhanced treatment effect. A well-known risk in subgroup analysis is that of false positives [Yusuf et al., 1991, Peto et al., 1995, Assmann et al., 2000, Brookes et al., 2001, Cui et al., 2002, Pocock et al., 2002, Brookes et al., 2004, Rothwell, 2005, Lagakos, 2006, Wang et al., 2007]. One way to reduce this risk is to pre-define a small number of potential subgroups before looking at the data; however, one may not always know *a priori* which subgroups may be of interest. When considering pre-defined subgroups, one common approach to reducing the risk of false positive findings is to consider a multiple testing procedure, such as a Bonferroni correction. Multiple testing procedures can effectively reduce false positives, but generally also lack power to detect true subgroups. Thus, one may instead wish to pre-define a statistical approach for identifying subgroups [Ruberg et al., 2010], which can be implemented using the data [Friedman and Fisher, 1999, Negassa et al., 2005, Su et al., 2008, 2009, Brinkley et al., 2010, Cai et al., 2011, Foster et al., 2011, Lipkovich et al., 2011, Qian and Murphy, 2011, Zhao et al., 2011, Imai and Ratkovic, 2012, Zhang et al., 2012]. Such pre-defined approaches can be very effective, but may still be prone to identifying subgroups, even when no true underlying subgroup exists [Foster et al., 2011]. To help quantify the potential usefulness of identified subgroups, Foster et al. [2011] considered the use of a metric, $Q(A)$, which is defined as the difference between the expected treatment effect in some subset of the covariate space, A, and the overall treatment effect. Such a metric can also be helpful for eliminating false subgroups, as small values suggest a lack of enhancement, but may be more effective if one can also obtain p-values for the metric estimates.

Before obtaining a p-value, one must consider what “null” means in a particular setting. Subgroups of enhanced (or more generally, different) treatment effect occur when the effect of treatment depends on the covariate values, i.e., such subgroups arise when treatment-by-covariate interactions exist. Thus, in this setting, “null” data may be defined as data in which no treatment-by-covariate interactions exist. Alternatively, if one has implemented a subgroup identification procedure, and wishes to evaluate the identified region of the covariate space, say A, null data could be defined as data for which the treatment effect in region A is equal to the overall treatment effect. In this paper, we will focus on the more general null scenario that no treatment-by-covariate interactions exist.

Consider the following general model:

$$y_i = h(\boldsymbol{x}_i) + g(\boldsymbol{x}_i)(T_i - \pi) + \epsilon_i, \tag{4.1}$$

where y_i is the observed, continuous outcome measure for subject i, $\boldsymbol{x}_i$ is the i^{th} row of the standardized, $n \times p$ covariate matrix $\boldsymbol{X}$, and T_i is a binary indicator of which treatment subject i received. Additionally, $\epsilon_i,\ i = 1, \ldots, n$ are iid errors with mean zero and variance σ^2 which are independent of the covariates, and π is a treatment randomization probability. We wish to consider a general class of models, so h and g are unspecified. Note that in this model, $g(\boldsymbol{x}_i)$ is the treatment effect at covariate value $\boldsymbol{x}_i$. Our goal is to test the null hypothesis that $g(\boldsymbol{x})$ is constant with respect to the covariates $\boldsymbol{x}$. That is, we wish to test for treatment-by-covariate interactions. One possible approach is to develop a test statistic for which asymptotic properties can be obtained, but this may be difficult for some methods. In such cases, a common approach is to use permutation tests.

A number of authors, including Edgington [1986], Good [2000], Potthoff et al. [2001], Bůžková et al. [2011] and Simon and Tibshirani [2012], have considered the testing of interactions using permutation-based methods, which work by shuffling parts of the data, such as the outcome, some or all of the covariates, or the treatment indicators, with the goal of eliminating one or more specific associations. In this chapter, we have a broad definition of what a permuted data set is. In general, permutations shuffle selected parts of the data set in such a way that the new data "look like" the original data in some aspects, but in other aspectes, the new data differ from the original data. For example, marginal distributions are preserved, but some associations between selected variables are not preserved. Moreover, with simple permutations, it may not be possible to completely control which aspects of the original data are preserved, and which are changed. An issue which must be addressed if one wishes to test for interactions using permutation-based methods is that it is generally impossible to remove only the associations of interest by simply permuting the data [Edgington, 1986, Good, 2000, Potthoff et al., 2001, Bůžková et al., 2011, Simon and Tibshirani, 2012]. One way to reduce the drawbacks of removing more than just the association of interest is to “cleverly” choose a test statistic [Edgington, 1986, Good, 2000, Potthoff et al., 2001, Bůžková et al., 2011], but there may not always be an obvious choice. The “permutation-like” methods that we will develop and describe later in this chapter have a similar characteristic of preserving some aspects of the data structure, while not preserving other aspects, but in a somewhat more controlled fashion. In particular, we will propose alternative forms of permutation methods, which are designed to remove only the associations of interest, thereby avoiding some of the potential issues with using traditional permutation tests for interactions.

The remainder of this chapter is as follows. In Section 4.2 we briefly review permutation tests and present a variety of alternative methods for obtaining p-values. In Section 4.3, the proposed methods are compared with a number of commonly-used permutation-based approaches in a simulation study. In Section 4.4, the proposed methods are implemented on real data from a randomized clinical trial, and in Section 4.5 we present a discussion.

4.2 Permutation tests

4.2.1 Review of permutation tests

Suppose we have observed data $(y_1, x_1), \ldots, (y_n, x_n)$, where y is some outcome and x is a covariate of interest, and that we only consider testing the null hypothesis of no association between y and x. This is generally done by calculating a test statistic, say $\widehat{TS}$, and then obtaining a p-value based on the null distribution of $\widehat{TS}$. In lieu of using asymptotics to determine the null distribution of $\widehat{TS}$, one may wish to employ a permutation test. For testing the null hypothesis that no association exists between x and y, approximately “null” data can be obtained by permuting either x or y. This is essentially the same as replacing the permuted covariate with a randomly generated noise covariate of the same distribution. Permutation tests work by repeating the process many times (say K), each time calculating a new value of the test statistic, $\widehat{TS}^{(k)}$. These K values are then used to define an approximated null distribution for $\widehat{TS}$, from which a p-value can be obtained.

As noted by Edgington [1986], there are two basic methods of permuting data. One approach is the *systematic* [Edgington, 1986] method, in which all possible permutations are considered; however, in this case, $K = n!$ (if the observed values are unique), so for even moderate sample sizes, the number of possible permutations will be very large. An alternative approach is *random* [Edgington, 1986] or Monte Carlo permutation, in which only a random subset of the possible permutations are considered, making it much more feasible for moderate-to-large data sets. In this paper, we consider only random permutation tests.

In the next subsection, we consider several methods which are permutation based, but are specifically designed to test only for interactions, which is something traditional permutation tests are generally unable to do. Though the proposed methods could be applied to any interaction, we will limit our discussion to treatment-by-covariate interactions.

4.2.2 Modified permutation methods

Permutation tests may not be ideal if one wishes to test for interactions, as they will generally be unable to remove only the association of interest. To better understand why this is, consider the following example. Suppose we have RCT data $(y_i, T_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, and that we wish to fit the interaction model $y_i = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}_i + \gamma(T_i - \pi) + \boldsymbol{\theta}^T\boldsymbol{x}_i(T_i - \pi) + \epsilon_i$, where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are $p \times 1$. In particular, suppose we wish to test for an interaction between treatment and one covariate, say x_j. That is, we wish to test the null hypothesis $\theta_j = 0$. Permuting y will eliminate the desired interaction, but will also eliminate all main effects of the covariates and the treatment, thereby making the "true" underlying model $y_i = \beta_0 + \epsilon'_i$. Similarly, permuting the covariate x_j or the treatment indicator T will eliminate the interaction, but also the corresponding main effect for either x_j or T, leading to a "true" underlying model where $\beta_j = \theta_j = 0$ or $\gamma = 0, \boldsymbol{\theta} = \mathbf{0}$ respectively.

One simple way to address this issue is considered by Potthoff et al. [2001], and involves permuting the covariates of interest within levels of T. Thus, in this case, one would permute values of x_j separately for those with $T_i = 1$ and those with $T_i = 0$. This approach will avoid removing the main effect for treatment, but still eliminates the main effect for x_j, so that the null model has $\beta_j = \theta_j = 0$. Note that if one wishes to test for all x-by-T interactions (null hypothesis $\boldsymbol{\theta} = \mathbf{0}$), this approach is equivalent to permuting y within levels of T.

Consider now the general model (4.1), and suppose we are interested in testing whether or not the treatment effect, $g(\boldsymbol{x}_i)$, is constant with respect to the covariates $\boldsymbol{x}_i$. Additionally, suppose that model (4.1) has been fit, giving function estimates $\hat{h}$ and $\hat{g}$, and let $\widehat{TS}$ be the sample value of the test statistic of interest. Our methods are designed to perturb the values of $\hat{g}(\boldsymbol{x}_i)$ to obtain a new treatment effect, g_i^*, which does not depend on the covariates, and which will be discussed later. Using this null treatment effect, along with the specific form of model (4.1), we obtain p-values as follows:

1. Create a new "null" outcome, $y_i^{*(k)} = \hat{h}(\boldsymbol{x}_i) + g_i^*(T_i - \pi) + \tilde{e}_i^{(k)}, i = 1, \ldots, n$, where $\tilde{e}_1^{(k)}, \ldots, \tilde{e}_n^{(k)}$ are randomly sampled (without replacement) centered residuals

from fitted model (4.1).

2. Using $\boldsymbol{y}^{*(k)}$, $\boldsymbol{X}$ and $\boldsymbol{T}$, refit model (4.1), and obtain a new value of the test statistic, $\widehat{TS}^{(k)}$.

3. Repeat steps 1 and 2 K times (i.e. $k = 1, \ldots, K$), giving K values of the test statistic, and use these values to obtain an approximate null distribution for the test statistic.

4. Use this approximate null distribution and the test statistic value from the observed data, $\widehat{TS}$, to obtain a p-value (either one or two-sided).

We now consider methods which use the estimated treatment effect, $\hat{g}$, to obtain the null treatment effect, g^*.

Mean of estimated treatment effect:

Fixed g^* approach. Perhaps the simplest null scenario in this setting is that of a constant treatment effect for all individuals. Thus, one natural choice is to create null data by giving all individuals the average estimated treatment effect, so that $g_i^* = \frac{1}{n}\sum_{i=1}^{n} \hat{g}(\boldsymbol{x}_i) \equiv \bar{\hat{g}}$, $i = 1, \ldots, n$. For the remainder of this paper, this will be referred to as the *fixed* g^* approach.

Fixed g^* (RN) approach. One could also consider a variation of the fixed g^* case, in which individuals have treatment effects which vary randomly around a fixed, population-wide mean. In this case, we define the null treatment effect for subject i to be $g_i^* = \bar{\hat{g}} + \tilde{\epsilon}_{g,i}$, where $\tilde{\boldsymbol{\epsilon}}_g$ is a random permutation of the non-centered residuals $\hat{\boldsymbol{g}} - \bar{\hat{g}}$. This will be refered to as the *fixed* g^* *random noise*, or *fixed* g^* (RN) approach.

Fixed g^* (BN) approach. Alternatively, if one has reason to believe these residuals do not vary equally about $\bar{\hat{g}}$ for all subjects, one may instead consider $g_i^* = g_{i,Bern}^* = \bar{\hat{g}} + \epsilon_{g,Bern,i}$, where

$$\epsilon_{g,Bern,i} = \begin{cases} \hat{g}_i - \bar{\hat{g}} & \text{when } a = 1 \\ \\ \bar{\hat{g}} - \hat{g}_i & \text{when } a = 0 \end{cases},$$

and a is an independently generated Bernoulli($\frac{1}{2}$) random variable. This will be referred to as the *fixed g^* Bernoulli noise*, or *fixed g^** (BN) approach.

Randomly shuffled estimated treatment effect:

Random g^* approach. As an alternative to the fixed treatment effect case, one may wish to consider a scenario in which each subject's response to treatment is random, but does not depend on the covariates. Therefore, we could also consider creating null data by giving each individual a random estimated treatment effect, so that $g_i^* = \tilde{g}_i$, where $\tilde{\boldsymbol{g}}$ is a random permuation of the estimated treatment effects. Note that this method is actually identical to the fixed g^* (RN) approach. Thus, we will not consider them separately.

Random g^* (RN) approach. As with the *fixed g^** approach, one could also consider the addition of random noise to the random treatment effects, $\tilde{\boldsymbol{g}}$. This could be done by following an approach similar to that used in the fixed g^* method. Specifically, we may consider $g_i^* = \tilde{g}_i + \tilde{\epsilon}_{g,i}$, where $\tilde{\boldsymbol{\epsilon}}_g$ is as defined above. This will be referred to as the *random g^* random noise*, or *random g^** (RN) approach.

Random g^* (BN) approach. Alternatively, we may consider $g_i^* = \tilde{g}_{i,Bern}^*$, where $\tilde{\boldsymbol{g}}_{Bern}^*$ is a random permutation of $\boldsymbol{g}_{Bern}^*$ (defined above). This will be referred to as the *random* g^* *Bernoulli noise*, or *random* g^* (BN) approach.

Constrained least-squares approach. One may also consider obtaining null data by modifying the estimated treatment effects so that they are approximately null, but with similar values to the original estimates. To do this, we use least squares to calculate a new treatment effect, g_i^*, which is close to $\hat{g}(\boldsymbol{x}_i)$, but has marginal sample correlations of zero with all the covariates. That is, we choose values g_i^*, $i = 1, \ldots, n$ which minimize

$$\sum_{i=1}^{n} \{\hat{g}(\boldsymbol{x}_i) - g_i^*\}^2$$

under the constraint that the sample correlations between $\boldsymbol{g}^*$ and $\boldsymbol{x}_j$, $j = 1, \ldots, p$, are zero, or equivalently by minimizing

$$\sum_{i=1}^{n} \{\hat{g}(\boldsymbol{x}_i) - g_i^*\}^2 + \sum_{j=1}^{p} \lambda_j \sum_{i=1}^{n} g_i^* x_{ij}$$

with respect to g^*, where λ_j, $j = 1, \ldots, p$, are Lagrange multipliers. Note that if the covariates are not centered, the penalty term becomes $\sum_{j=1}^{p} \lambda_j \sum_{i=1}^{n} (g_i^* - \bar{g}^*)(x_{ij} - \bar{x}_j)$. It is straightforward to show that

$$\boldsymbol{g}^* = \hat{\boldsymbol{g}} - \frac{1}{2}\boldsymbol{X}\boldsymbol{\lambda}, \qquad (4.2)$$

$$\text{where } \boldsymbol{\lambda} = 2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{g}}.$$

This method will be referred to as the *Lagrange* g^* approach. It should be noted that a correlation of zero between the treatment effect and a covariate does not necessarily mean the treatment effect is independent of the covariate. Though we don't discuss it here, one could also consider additional constraints, such as g^* being uncorrelated with x_j^2, or g^* being uncorrelated with $\hat{h}$.

To better understand why the Lagrange method should work, note that $\boldsymbol{g}^*$ is exactly equal to the residuals from the model $\hat{g}(\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\omega} + \epsilon_i$. Thus, the Lagrange method works by estimating the contribution of the covariates to the treatment effect, and then removing this estimated contribution, so that only that part of the estimated treatment effect which does not depend on the covariates remains.

4.3 Simulations

To evaluate the performance of the proposed methods under a variety of scenarios, a simulation study was performed. In addition to the proposed methods, we considered four permutation-based methods:

- Permutation of y;
- Permutation of $\boldsymbol{X}$;
- Permutation of T;
- Permutation of y within levels of T.

These approaches will be referred to as the *permute* y, *permute* $\boldsymbol{X}$, *permute* T and *permute* y (in T) methods respectively. We begin by considering a linear model with treatment-by-covariate interactions, which allows us to compare the proposed methods to both permutation and exact p-values. We then discuss simulation results for a more complex model.

4.3.1 Simple linear model example

Data were generated from the model

$$y = 3 + \beta\Big(x_1 + x_3 + x_5\Big) + \gamma\Big(T - \frac{1}{2}\Big) + \theta\Big(T - \frac{1}{2}\Big)\Big(x_1 + x_2\Big) + \epsilon,$$

where ϵ's and all x's are iid standard normal, and the design matrix, $\boldsymbol{X}$, has five columns. In this case, the test of interest (for the assumed model, given below) is that $\theta_j = 0,\ j = 1, \ldots, 5$. This is a situation where the F statistic has a known distribution, and thus an exact test exists. We considered four scenarios:

(1) $\theta = 0.35$, $\gamma = 0.25$, $\beta = 0.5$;

(2) $\theta = 0$, $\gamma = 0.25$, $\beta = 0.5$;

(3) $\theta = 0.35$, $\gamma = 0.5$, $\beta = 0.25$;

(4) $\theta = 0$, $\gamma = 0.5$, $\beta = 0.25$.

In scenarios (1) and (2), the main effects for the covariates are large, and in scenarios (3) and (4) the main effect for treatment is large. These cases were chosen to illustrate the potential shortcomings of traditional permutation-based tests when large main effects exist. In each scenario, 500 data sets of size 200 were generated, and the model $y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + \gamma(T - \frac{1}{2}) + \boldsymbol{\theta}^T\boldsymbol{x}(T - \frac{1}{2}) + \epsilon$ was fit, where $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, are 5×1 vectors. The null hypothesis was that no x-by-T interactions exist (vs. alternative that at least one exists). To assess the sensitivity of the methods to the choice of test statistic, we computed the permutation p-values using an F statistic and also the statistic $\sum \hat{\theta}_j^2$. For all non-asymptotic methods, 1000 permutations were used (so $K = 1000$).

Note that, in this case, the covariates are independent, as are the observations, and the design matrix is standardized. As a result, it can be shown that $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is approximately equal to $(\frac{p}{n})\boldsymbol{I}$, where $\boldsymbol{I}$ is an $n \times n$ identity matrix. For the fitted model, the estimated treatment effect vector $\hat{\boldsymbol{g}}$ is equal to $\hat{\boldsymbol{\gamma}} + \boldsymbol{X}\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}, \ldots, \hat{\gamma})^T$. Thus for the Lagrange method, from (4.2) we have

$$\begin{aligned}
\boldsymbol{g}^* &= \hat{\boldsymbol{g}} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{g}} \\
&= \hat{\gamma} + \boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\gamma} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\theta}} \\
&= \hat{\gamma} + \boldsymbol{X}\hat{\boldsymbol{\theta}} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\gamma} - \boldsymbol{X}\hat{\boldsymbol{\theta}} \\
&= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\hat{\gamma} \\
&\approx \left(1 - \frac{p}{n}\right)\hat{\gamma} \approx \hat{\boldsymbol{\gamma}},
\end{aligned}$$

where the last approximation is a result of the fact that, in this case, p is considerably smaller than n, so that $\frac{p}{n}$ is nearly zero. In addition, for the fixed g^* method, we have

$$g_i^* = \frac{1}{n}\sum_i(\hat{\gamma} + \boldsymbol{x}_i^T\hat{\boldsymbol{\theta}}) = \hat{\gamma} + \frac{1}{n}\sum_j\left(\hat{\theta}_j\sum_i x_{ij}\right) = \hat{\gamma},$$

again because covariates are standardized, so that the second term is exactly zero. Therefore, for this specific example, the Lagrange and fixed g^* methods are nearly identical. Note that if the covariates were correlated, the off-diagonal elements of $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ would be non-zero, leading to unique Lagrange g^* values for each subject. Thus, in this case the Lagrange and Fixed g^* methods would be different.

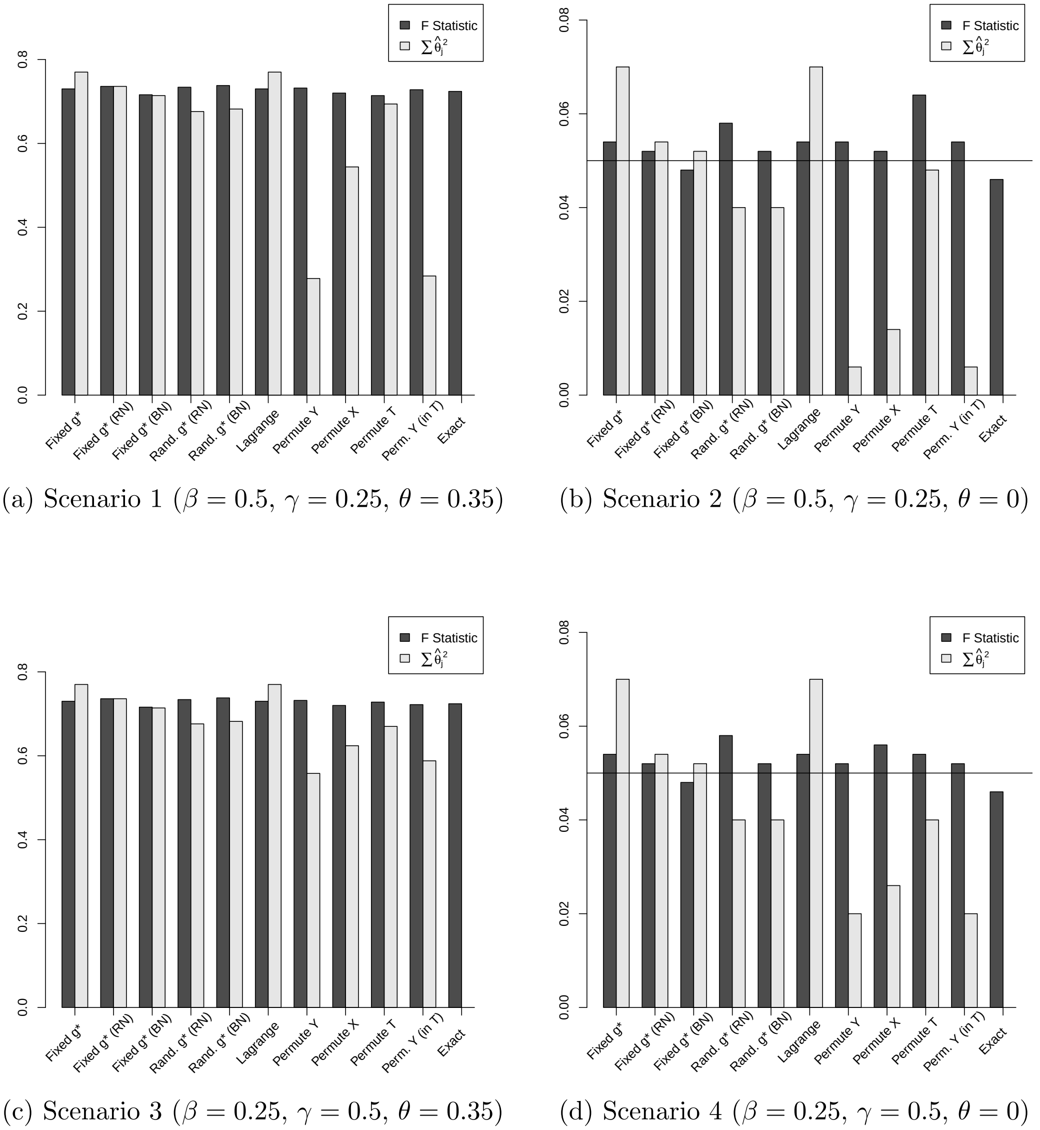
Looking at Figure 4.1, we can see that rejection rates when using the F statistic are generally quite similar for all methods in all four scenarios, whereas the rejection rates for the statistic $\sum\hat{\theta}_j^2$ vary noticeably between methods. In particular, the $\sum\hat{\theta}_j^2$

rejection rates for the permute $\boldsymbol{X}$, permute y and permute y (in T) methods tend to be considerably lower than those for the other methods. This is due to the fact that these methods remove several main effects (in addition to the desired interactions) in the process of creating new “null” outcome values, which subsequently have larger error variances than the observed outcome values. Because $\sum \hat{\theta}_j^2$ comes from the parameter estimates for a linear regression model, its variance depends on the variance of the outcome values, and in particular will increase or decrease as the error variance of the outcome increases or decreases. Thus, the permute $\boldsymbol{X}$, permute y and permute y (in T) methods induce “null” distributions for $\sum \hat{\theta}_j^2$ which have much larger variances than the correct null distribution (where only the interactions are removed), leading to larger p-values and fewer rejections. A similar effect can be seen in the permute T method, but to a lesser degree, as this method only removes one main effect. In contrast, the Lagrange and Fixed g^* methods create “null” outcome values whose variances are slightly smaller than that of the observed outcome values, causing these methods to give elevated rejection rates. This general phenomenon is mentioned by Bůžková et al. [2011], who note that permutation-based methods perform better for statistics with a pivotal null distribution.

Overall, the best methods (not based on the exact F distribution) are fixed g^* (BN) and fixed g^* (RN), which have rejection rates that are nearly identical for both the F statistic and $\sum \hat{\theta}_j^2$, and which are very close to 0.05 in the “null” scenarios (2 and 4). The random g^* (BN) and random g^* (RN) methods are also fairly good, but have slightly lower rejection rates for $\sum \hat{\theta}_j^2$, again due to an elevated outcome error variance (from adding “noise” twice). Figures 4.2-4.5 further illustrate the effect of creating “null” data which has an outcome error variance that is too large (or too small). This is again most noticeably for the permute $\boldsymbol{X}$, permute y and permute y

(in T), methods, whose histograms have heavier right tails for scenarios 1 and 3 and are obviously non-uniform in scenarios 2 and 4.

The results for the statistic $\sum \hat{\theta}_j^2$ help illustrate the importance of an appropriately chosen permutation method when the test statistic is somewhat ad hoc, as may be the case in situations where asymptotics are difficult to obtain.



(a) Scenario 1 ($\beta = 0.5$, $\gamma = 0.25$, $\theta = 0.35$)

(b) Scenario 2 ($\beta = 0.5$, $\gamma = 0.25$, $\theta = 0$)

(c) Scenario 3 ($\beta = 0.25$, $\gamma = 0.5$, $\theta = 0.35$)

(d) Scenario 4 ($\beta = 0.25$, $\gamma = 0.5$, $\theta = 0$)

Figure 4.1: Rejection rates ($\alpha = 0.05$) for simple linear model simulations

4.3.2 Complex model example

As mentioned in the Introduction, the proposed methods were motivated by our interest in subgroup analysis. Thus, we also considered p-values for the metric $Q(\hat{A})$ for subgroups identified using the Average Value approach proposed in Chapter 2. As noted by Foster et al. [2011], $Q(\hat{A})$ can be viewed as the treatment effect enhancement in the region $\hat{A}$ beyond the overall treatment effect. In this case, we are interested in testing whether or not the identified subgroup is "real," so the null hypothesis is that $Q(\hat{A})$ is 0 (vs. alternative that $Q(\hat{A}) > 0$). We considered four scenarios for the true treatment effect, g:

(1) $g(\boldsymbol{x}) = 15I(x_1 > 0, x_2 > 0)$;

(2) $g(\boldsymbol{x}) = 5$ (case 5 from Chapter 2);

(3) $g(\boldsymbol{x}_i) = 5 + e_i$, $i = 1, \ldots, n$, where e_i's are iid N(0,5^2);

(4) $g(\boldsymbol{x}) = 0$ (case 6 from Chapter 2).

Other than the treatment effect, the data generation model was identical to that in the Chapter 2 simulations. That is, data were generated from the model $y_i = 30 + 5x_{1i} + 5x_{2i} - 5x_{7i} + T_i g(\boldsymbol{x}_i) + \epsilon_i$, where x's are iid standard normal and ϵ's are iid normal with mean zero and variance 100 and are independent of the x's. Scenario 1 is a modified version of case 1 from Chapter 2, with the coefficient of g reduced from 35 to 15 to reduce the power to detect a positive $Q(\hat{A})$ (which is near 0.9 when a coefficient of 35 is used). Scenario 3 can be viewed as a more challenging "null" case, where a main effect for treatment exists, but varies randomly around a population mean, so that some subjects do have an enhanced treatment effect, but enhancement is independent of the covariates. Alternatively, scenario 3 can be viewed as being like scenario 2, but with a larger error variance. For each scenario, 200 data sets of size 500 were generated.

We consider the null distributions of three statistics, $\hat{Q}(\hat{A})_{RS}$, $\hat{Q}(\hat{A})_{SND}$ and $\hat{Q}(\hat{A})_{\hat{g}}$, which are estimates of the enhancement metric $Q(\hat{A})$. The first statistic, $\hat{Q}(\hat{A})_{RS}$, is obtained by subtracting the observed marginal treatment effect, $\frac{\sum_i y_i T_i}{\sum_i T_i} - \frac{\sum_i y_i(1-T_i)}{\sum_i (1-T_i)}$, from the observed treatment effect in the region $\hat{A}$, $\frac{\sum_i y_i T_i I(\boldsymbol{x}_i \in \hat{A})}{\sum_i T_i I(\boldsymbol{x}_i \in \hat{A})} - \frac{\sum_i y_i(1-T_i) I(\boldsymbol{x}_i \in \hat{A})}{\sum_i (1-T_i) I(\boldsymbol{x}_i \in \hat{A})}$. This is expected to be positively biased, as the same data which were used to identify the region $\hat{A}$ are being used to estimate $Q(\hat{A})$. Note that independent outcome values from approximately the same distribution could be obtained by adding random residuals to the outcome estimates, $\hat{\boldsymbol{y}}$. Thus, one way to reduce the bias of $\hat{Q}(\hat{A})_{RS}$ is to obtain independent outcome measures as described above, and use these instead of y_i's to compute $\hat{Q}(\hat{A})_{RS}$. This could be repeated several times, and the resulting estimates could be averaged to obtain a less biased estimate of $Q(\hat{A})$; however, because the residuals have mean zero, this is approximately equivalent to just replacing the observed y values with $\hat{y}$s in $\hat{Q}(\hat{A})_{RS}$. Therefore, the second statistic, $\hat{Q}(\hat{A})_{SND}$, is obtained by computing $\hat{Q}(\hat{A})_{RS}$, but with y_is replaced by $\hat{y}_i$s. Because the Average Value method involves directly estimating the treatment effect, $Q(\hat{A})$ can also be estimated directly from these treatment effect estimates, $\hat{\boldsymbol{g}}$, by computing the difference $\frac{\sum_i \hat{g}(\boldsymbol{x}_i) I(\boldsymbol{x}_i \in \hat{A})}{\sum_i I(\boldsymbol{x}_i \in \hat{A})} - \frac{\sum_i \hat{g}(\boldsymbol{x}_i)}{n}$. This is the third statistic, and is referred to as $\hat{Q}(\hat{A})_{\hat{g}}$, and is very similar to $\hat{Q}(\hat{A})_{SND}$, with the two statistics being exactly equal in the case of paired data. To obtain p-values for these statistics, we again follow the procedure outlined in subsection 3.2.2. In this case, the Average Value procedure is re-implemented each time step 2 is performed, and new values of these three test statistics are calcuated. Because the Average Value approach is computationally expensive, p-values were computed using only 100 permutations (so $K = 100$). It is worth noting that the Average Value method involves nonparametric regression, and in this case we do not have an obvious test statistic, such as the F statistic in the simple linear model case, which has a known exact or asymptotic distribution for the fitted model (4.1).

From Figure 4.6, we can see that the statistic $\hat{Q}(\hat{A})_{\hat{g}}$ performs the best overall, showing generally high power for scenario 1, while also tending to be the closest to the desired type-I error rate of 0.05 (and rarely exceeding it) in the three null cases. The statistics $\hat{Q}(\hat{A})_{RS}$ and $\hat{Q}(\hat{A})_{SND}$ perform well for the permute $\boldsymbol{X}$ and permute y (in T) approaches in scenario 1, but nearly always noticeably exceed the desired type-I error level in the three null cases. The very high type-I errors for the RS and SND estimates observed in scenarios 2 and 3 for the permute y and permute T methods are most likely due to instability in these metrics when the identified region $\hat{A}$ is extremely small, as is often the case for the "null" data sets created by permuting y or permuting T.

Histograms of p-values for the four complex scenarios for all three metrics are given in Figures 4.7 - 4.10. For scenario 1 (Figure 4.7), the histograms generally have the desired "right-skewed" shape, with the exception of the permute y and permute T methods. For the null scenarios (Figures 4.8 - 4.10), we can see that histograms for the proposed methods are generally closer to uniform than those for the four simple permutation methods, especially for the statistic $\hat{Q}(\hat{A})_{\hat{g}}$. For the permute y and permute T methods, we frequently see a bimodal shape, which is again most likely a result of identifying very small regions $\hat{A}$ for "null" data created by permuting y or T. Though they showed generally good power and type-I error, we can see that p-values for the permute $\boldsymbol{X}$ and permute y (in T) methods are often noticeably non-uniform.

In general, the proposed methods perform very well, though the Lagrange method had slightly lower power in scenario 1 and generally much lower type-I errors in scenarios 2-4. The permute $\boldsymbol{X}$ and permute y (in T) methods also performed well in the four scenarios we considered, though we suspect the performance of these two methods would suffer more in non-null scenarios where g is more complex. Based on the results from the simple and complex model simulations, it appears that the fixed g^* and random g^* (and their variants with Bernoulli or random noise) are the best choice. Additionally, the results of the simulation study suggest that, if one wishes to obtain p-values for the metric $Q(\hat{A})$ for subgroups identified using the Average Value method, the best choice in test statistic is $\hat{Q}(\hat{A})_{\hat{g}}$.

As noted previously, the estimates $\hat{Q}(\hat{A})_{\hat{g}}$ and $\hat{Q}(\hat{A})_{SND}$ are quite similar, and in the case of paired data, identical. Thus, the noticeably performance differences between these two methods in our simulation study were surprising. We believe that these differences are a result of identifying subgroups which do not have a similar number of treated and control subjects. In particular, when very small subgroups are identified, these differences can be quite severe, and lead to very pronounced differences between the two estimates.

4.4 Application to data from a randomized clinical trial

Recall that in Chapter 2, the Average Value method procedure was applied to pre-hypertension data from the TROPHY study [Julius et al., 2006] in an attempt to identify the subset of individuals who should not receive candesartan, a treatment for hypertension. In this analysis, the identified region was $\hat{A} = \{\text{HDL:LDL ratio} < 0.38,\ \text{HDL cholesterol} < 46.02,\ \text{total insulin} \geq 25.11\}$, contained 20 subjects, and had corresponding (uncorrected) estimates of $Q(\hat{A})$ of -1.63, -8.14 and -9.75 for the RS, SND and Mean $\hat{g}$ methods respectively. To further assess the identified subgroup, the proposed methods and the four simple permutation methods were used to obtain p-values for these uncorrected estimates. Because the Average Value approach is computationally expensive, and because a finer grid of cutpoints was considered for the TROPHY data, further decreasing the computational speed of the method, p-values were computed using 100 permutations (so $K = 100$).

Table 4.1: P-values for TROPHY data

	$Q(\hat{A})$ Estimate		
Method	RS	SND	Mean $\hat{g}$
Fixed g^*	0.94	0.33	0.11
Fixed g^* (RN)	0.92	0.34	0.12
Fixed g^* (BN)	0.95	0.34	0.12
Random g^* (RN)	0.90	0.42	0.16
Random g^* (BN)	0.92	0.29	0.14
Lagrange	0.94	0.37	0.17
Permute y	0.46	0.00	0.00
Permute $\boldsymbol{X}$	0.96	0.25	0.08
Permute T	0.41	0.00	0.00
Permute y (in T)	0.90	0.21	0.04

P-values for these three estimates for the five proposed methods and the four simple permutation methods are given in Table 4.1. As might be expected, given the relative magnitudes of the three estimates, the Mean $\hat{g}$ statistic has the smallest p-value for all the methods considered, followed by the SND estimate, with the RS estimate having the largest p-values. Though somewhat large for the RS and SND statistics, the p-values for the Mean $\hat{g}$ were relatively small for the fixed g^* (RN) and fixed g^* (BN) approaches (which were shown to perform the best in our simulations). Thus, it is possible that a small subgroup of people who shouldn't receive candesartan exists, but can't be fully demonstrated with the current data.

When we permute y or T, the new treatment effect estimates are generally centered around zero, which in this case will most likely lead to a larger "don't treat" region than that identified using the observed data, for which nearly all treatment effect estimates were positive. These larger "don't treat" regions for the permuted data will generally have smaller corresponding estimates of $Q(\hat{A})$, which may be why the permute y and permute T approaches give the smallest p-values. Given the poor performance of the permute y and permute T approaches in our simulations, it may not be wise to trust these methods when testing for interactions, particularly when a somewhat ad hoc test statistic is being used.

4.5 Discussion

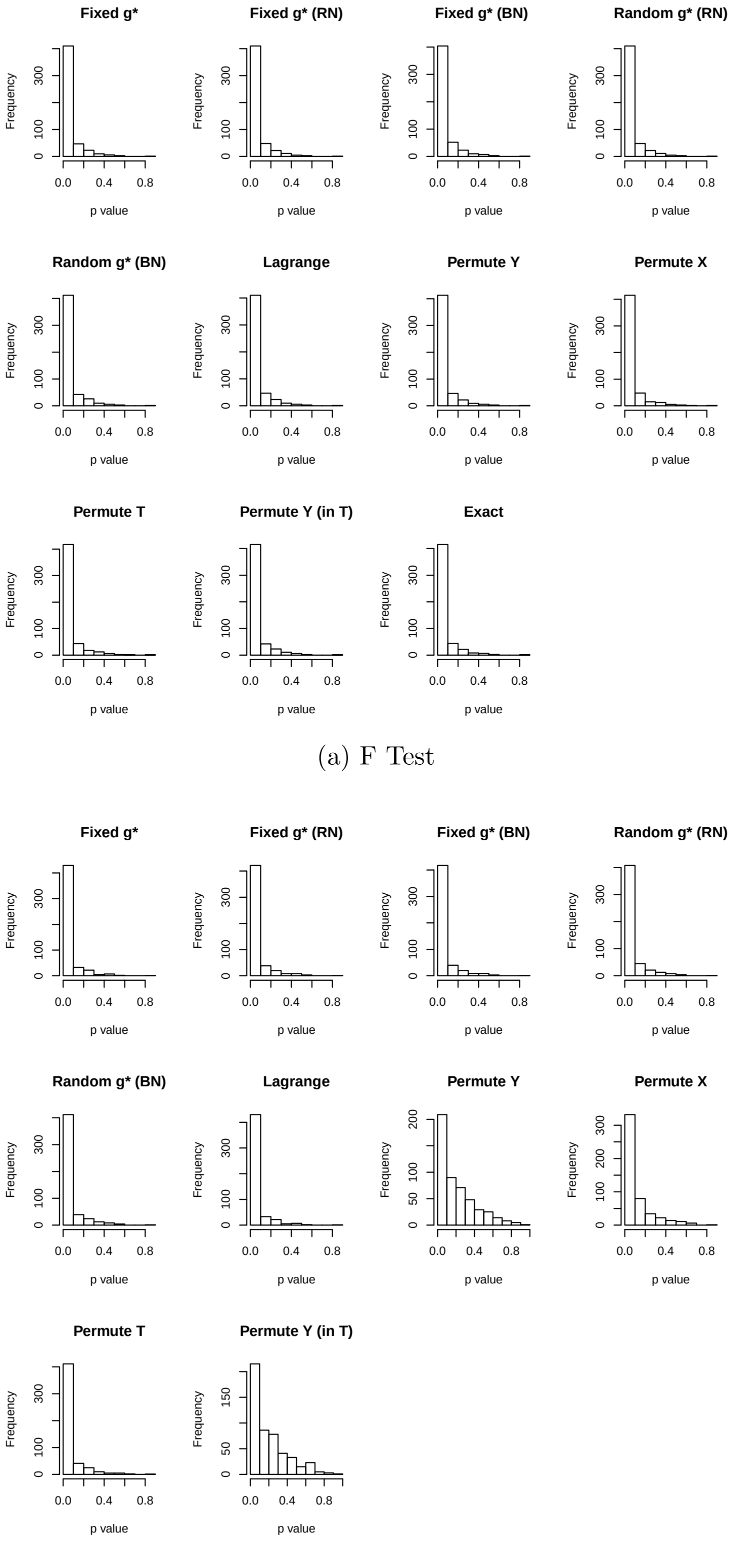
We proposed several permutation-based methods which can be used as an alternative to simple permutation test when one wishes to test for interactions. The proposed methods were shown to generally outperform simple permutation tests, particularly when we considered more complex scenarios and test statistics. These methods may help to reduce false positive findings when a pre-defined subgroup identification strategy such as Virtual Twins [Foster et al., 2011] or the Average Value procedure is employed.

In the example data analysis in this paper, we considered only the estimates $\hat{Q}(\hat{A})_{RS}$, $\hat{Q}(\hat{A})_{SND}$ and $\hat{Q}(\hat{A})_{\hat{g}}$ as test statistics for the Average Value procedure. In the future, it may be interesting to consider the performance of the proposed methods when alternative test statistics are used. One such alternative is the Z statistic for the test of interaction ($\alpha_3 = 0$) in the model $y = \alpha_0 + \alpha_1 T + \alpha_2 I(\boldsymbol{x} \in \hat{A}) + \alpha_3 T I(\boldsymbol{x} \in \hat{A}) + e$, which could be fit after the region $\hat{A}$ has been identified.

We show in the simple linear model simulations that the fixed g^* and Lagrange methods are slightly anti-conservative, and hypothesized that this may be due to the decreased total variance of y_i^* compared to y_i. Thus, for these methods, it may be helpful to consider inflating the variance of the residuals in step 1 of our algorithm, so that the variance of y_i^* matches that of y_i. Similarly, the random g^* (RN) and random g^* (BN) methods were overly conservative in our simple linear model simulations, which we believe is due to an increased total variance of y_i^* relative to y_i, so for these methods it may be helpful to consider deflating the variance of the residuals in step 1 of our algorithm.

We consider only a limited number of scenarios in our simulation study. Thus, in the future it may also be useful to consider a wider variety of simulation settings. Given our example randomized clinical trial data, it may be of interest to consider a scenario in which the treatment effects are mostly positive, so that our goal is to identify the small subset of patients who should not receive treatment. In this case, we could again consider setting the offset, δ, at zero. This may help us to better understand the results in our example data analysis. Additionally, it may be interesting to assess the performance of our methods for non-null cases in which the treatment effect is more complex, so that the form of potential subgroups considered in our search does not match the underlying truth.

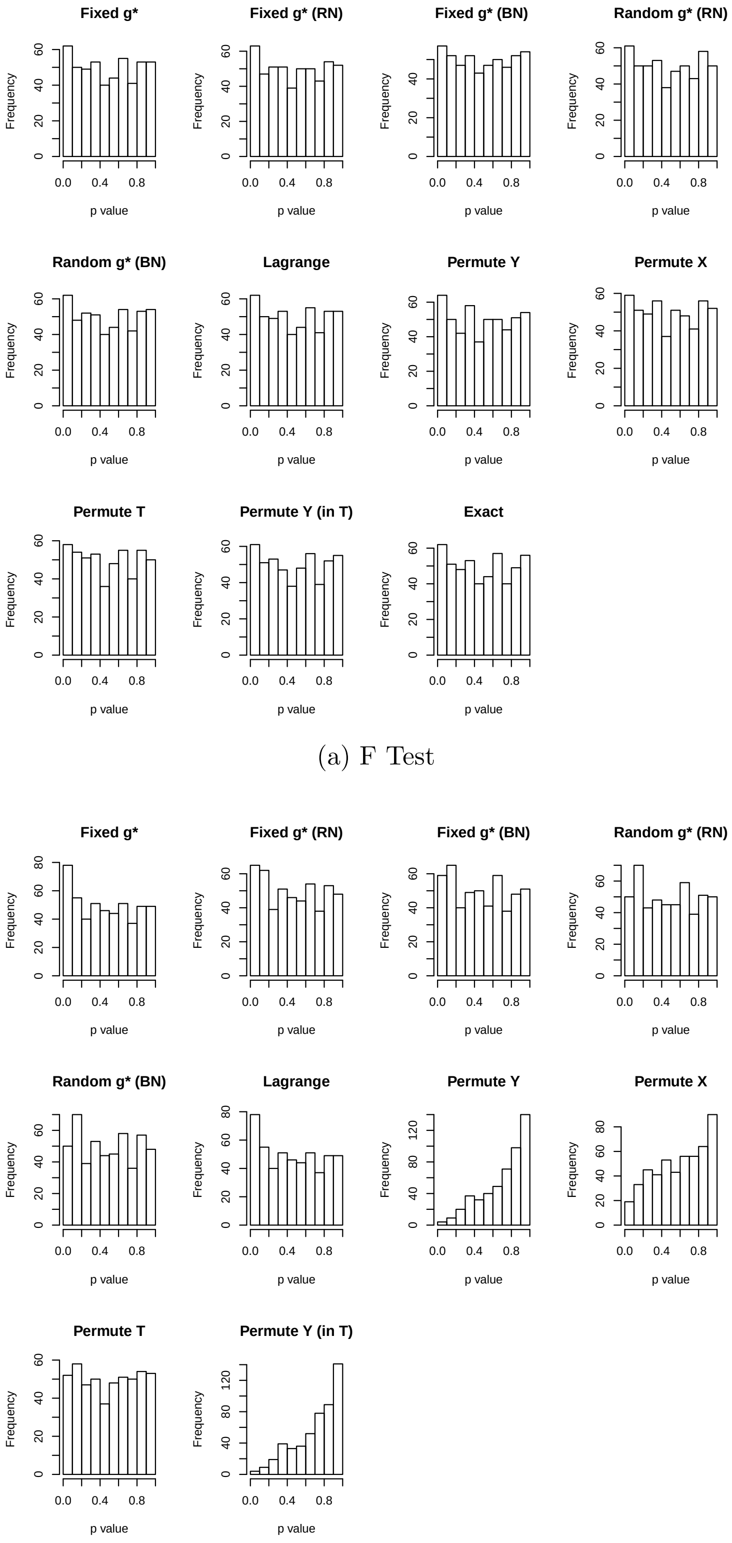
It may be possible to improve the performance of the various permutation methods by considering a number of refinements. For example, in the first step of our algorithm, we resample residuals from the fit of model 4.1 without replacement, but we could also consider resampling these residuals with replacement.



(a) F Test

(b) $\sum \hat{\theta}_j^2$

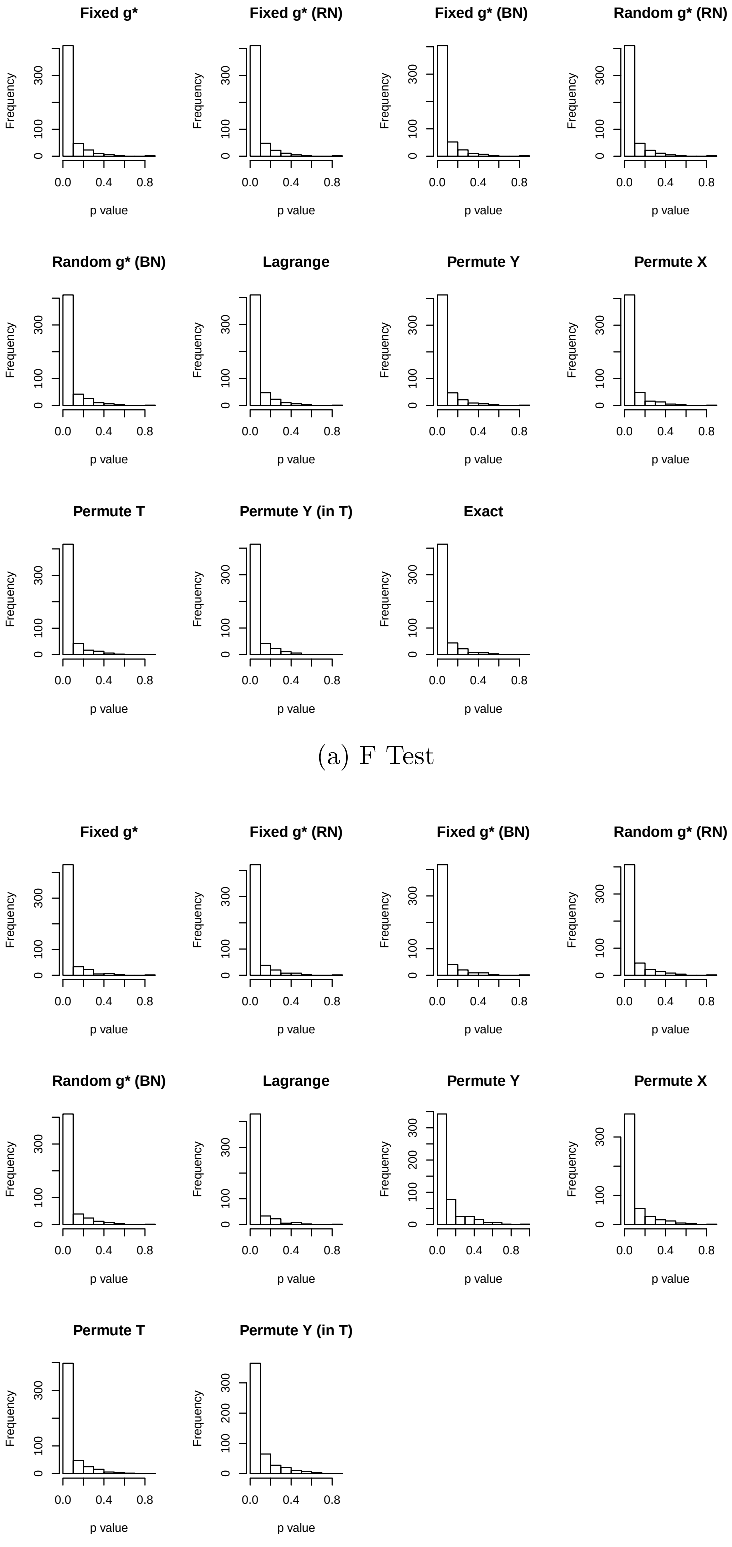
Figure 4.2: Histograms of p-values for simple linear scenario 1 ($\beta = 0.5$, $\gamma = 0.25$, $\theta = 0.35$)



(a) F Test

(b) $\sum \hat{\theta}_j^2$

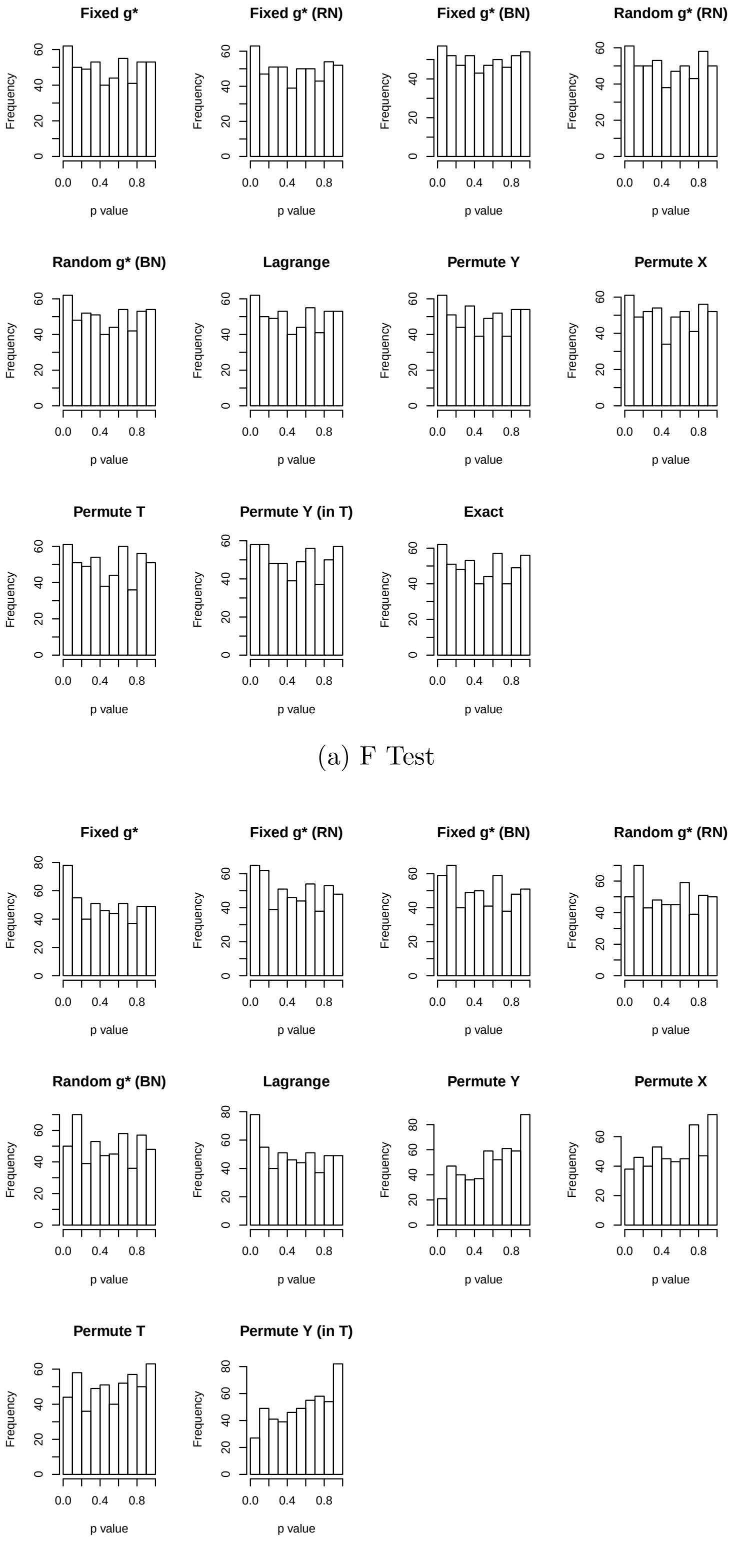
Figure 4.3: Histograms of p-values for simple linear scenario 2 ($\beta = 0.5$, $\gamma = 0.25$, $\theta = 0$)



(a) F Test

(b) $\sum \hat{\theta}_j^2$

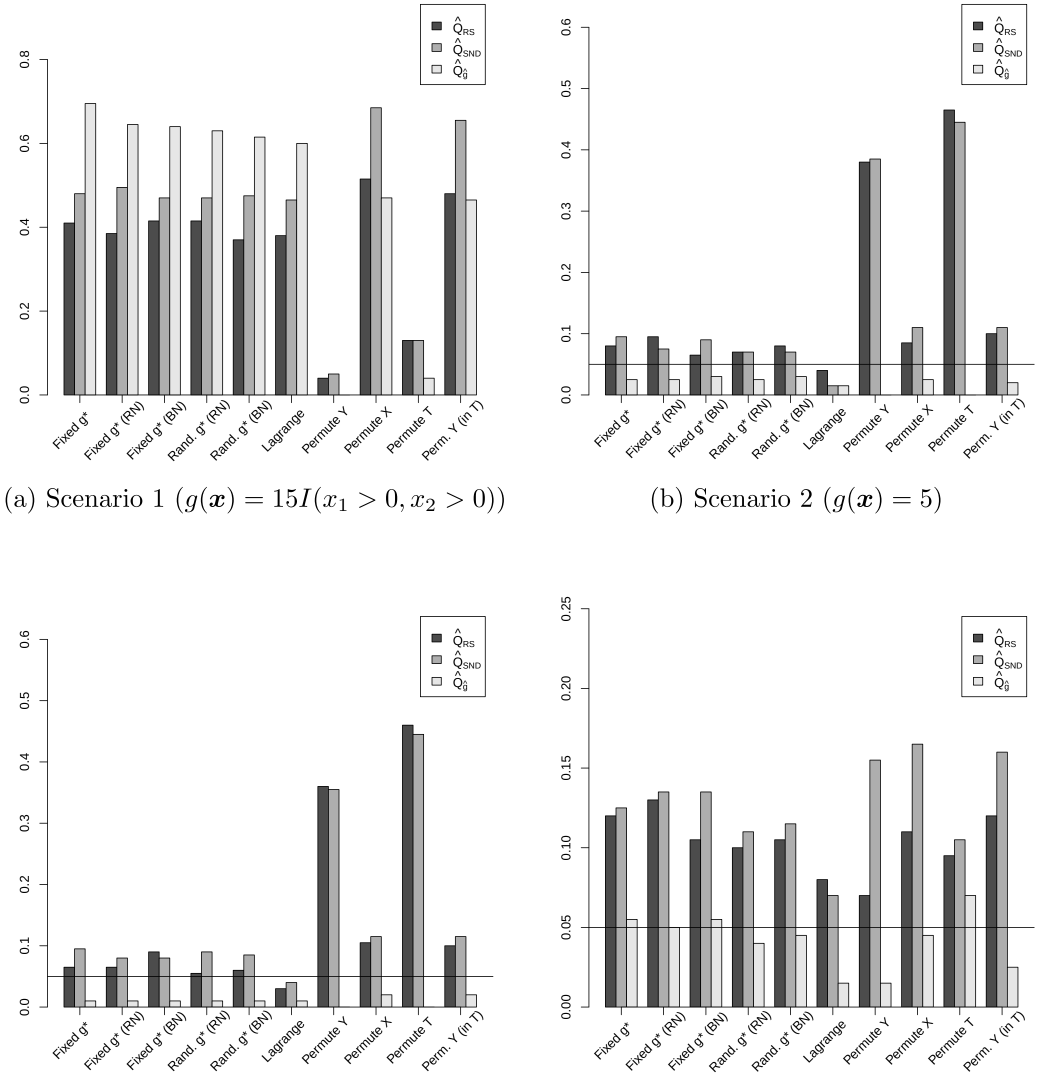
Figure 4.4: Histograms of p-values for simple linear scenario 3 ($\beta = 0.25$, $\gamma = 0.5$, $\theta = 0.35$)



(a) F Test

(b) $\sum \hat{\theta}_j^2$

Figure 4.5: Histograms of p-values for simple linear scenario 4 ($\beta = 0.25$, $\gamma = 0.5$, $\theta = 0$)



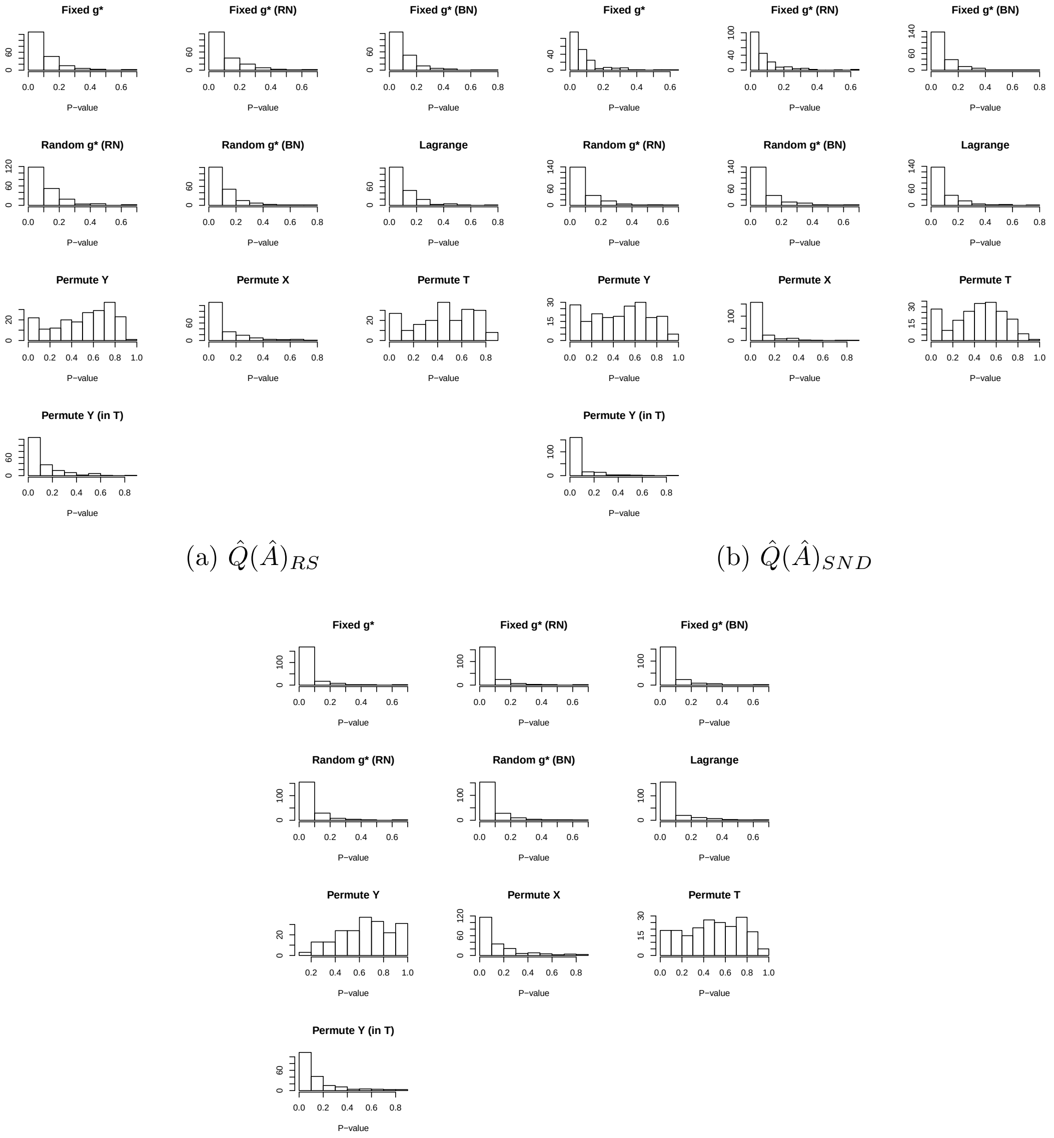
(a) Scenario 1 ($g(\boldsymbol{x}) = 15I(x_1 > 0, x_2 > 0)$)

(b) Scenario 2 ($g(\boldsymbol{x}) = 5$)

(c) Scenario 3 ($g(\boldsymbol{x}_i) = 5 + e_i$)

(d) Scenario 4 ($g(\boldsymbol{x}) = 0$)

Figure 4.6: Rejection rates ($\alpha = 0.05$) for complex model simulations

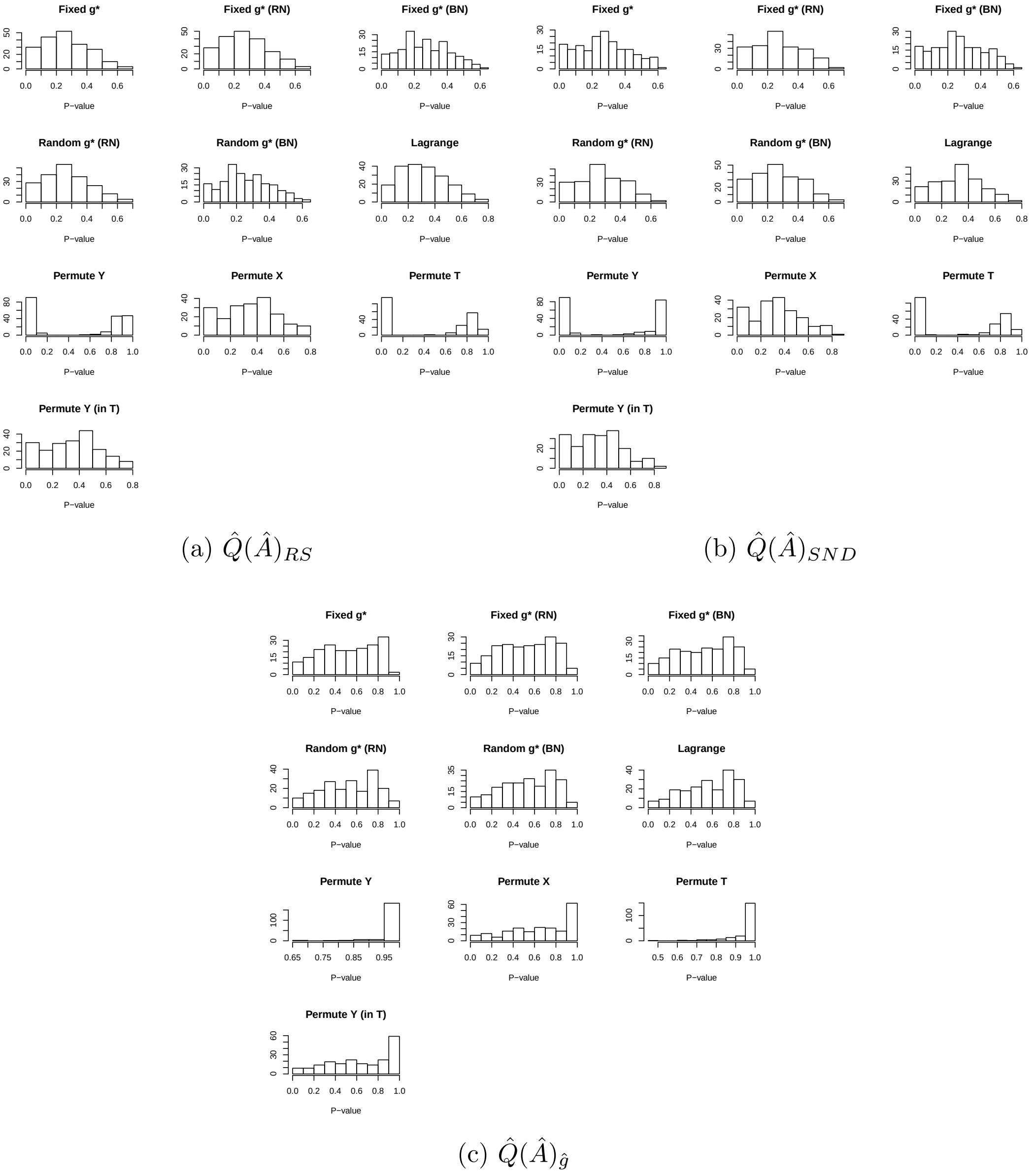


(a) $\hat{Q}(\hat{A})_{RS}$

(b) $\hat{Q}(\hat{A})_{SND}$

(c) $\hat{Q}(\hat{A})_{\hat{g}}$

Figure 4.7: Histograms of p-values for complex scenario 1 ($g(\boldsymbol{x}) = 15I(x_1 > 0, x_2 > 0)$)

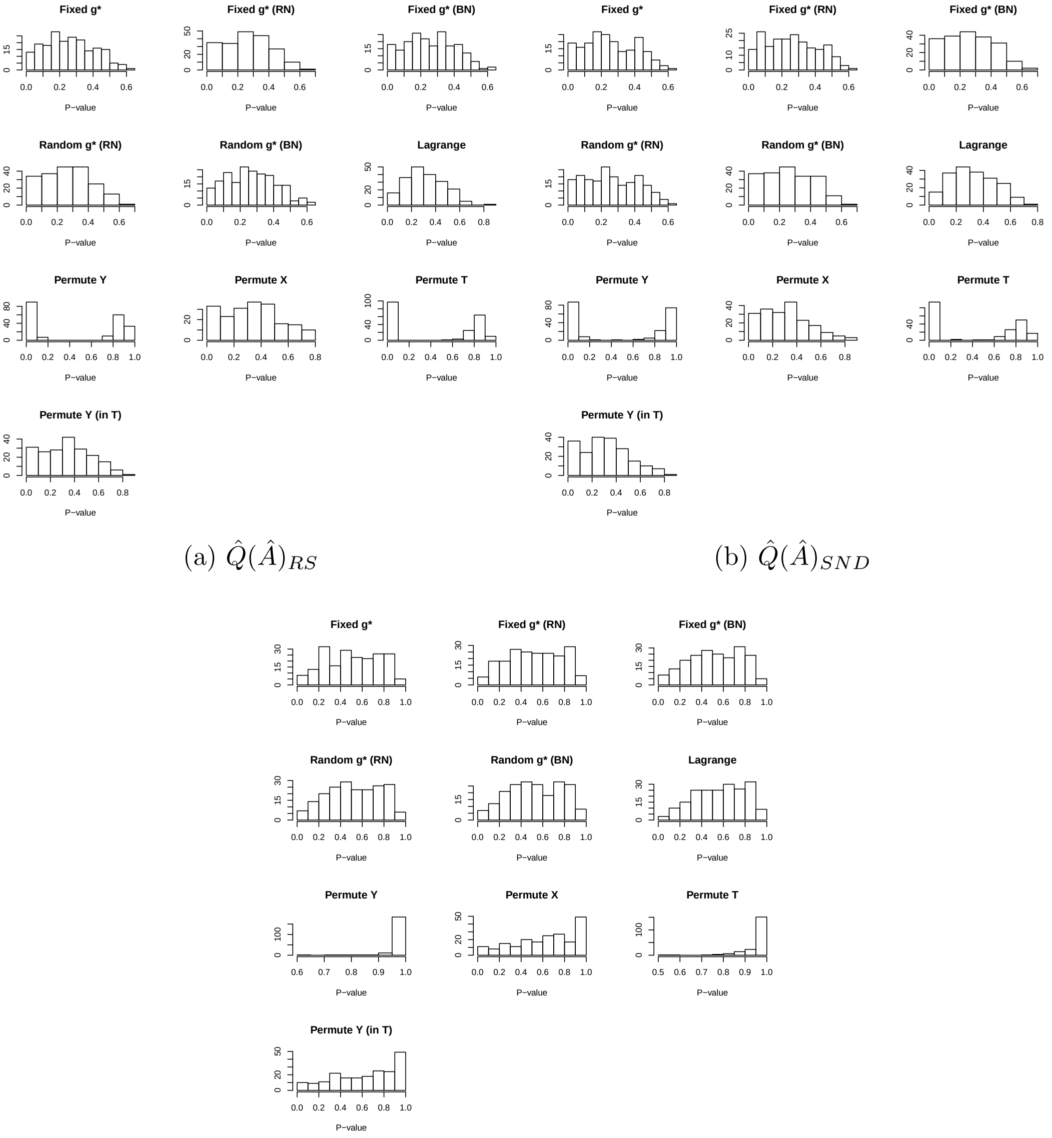


(a) $\hat{Q}(\hat{A})_{RS}$

(b) $\hat{Q}(\hat{A})_{SND}$

(c) $\hat{Q}(\hat{A})_{\hat{g}}$

Figure 4.8: Histograms of p-values for complex scenario 2 ($g(\boldsymbol{x}_i) = 5$)

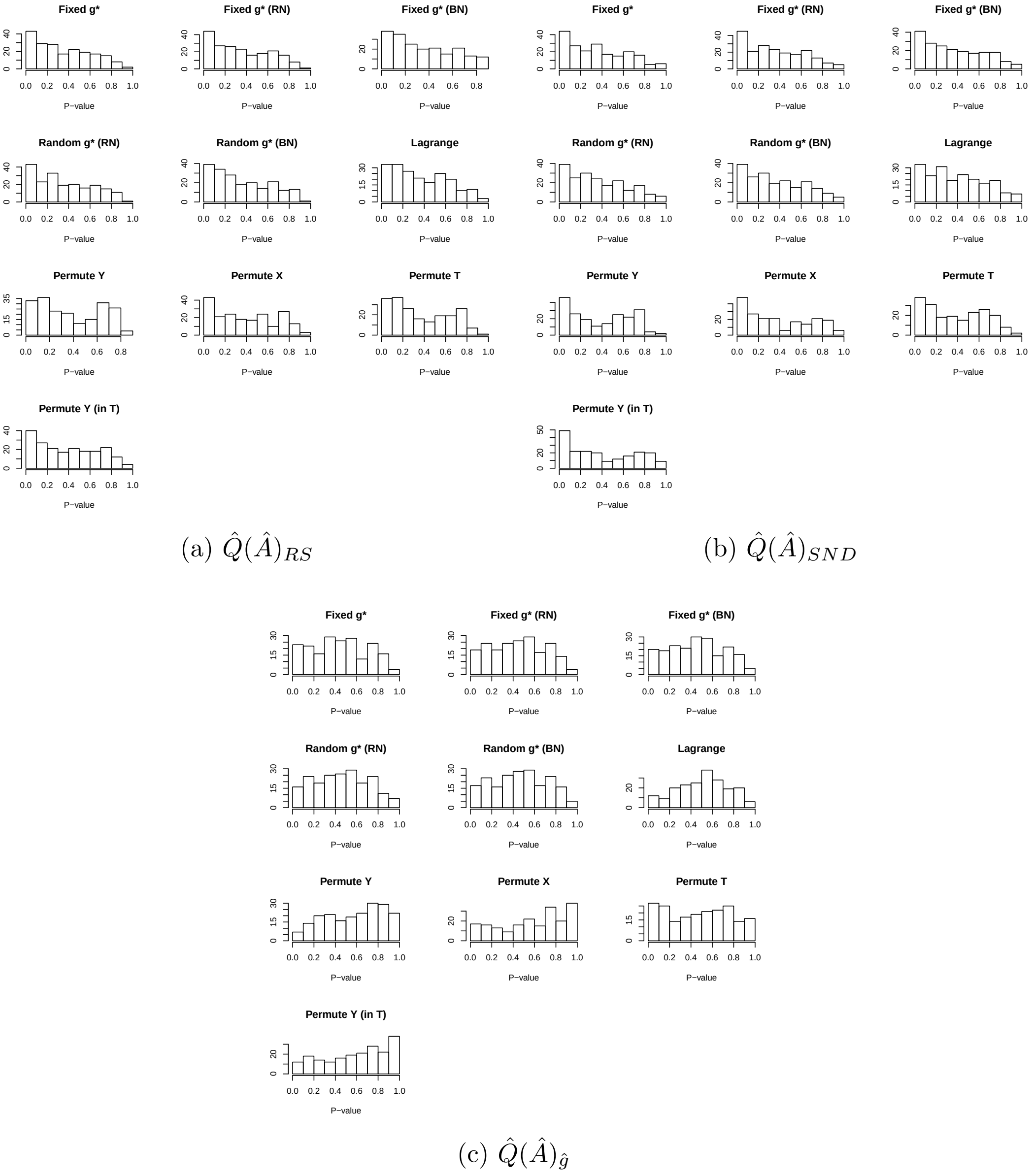


(a) $\hat{Q}(\hat{A})_{RS}$

(b) $\hat{Q}(\hat{A})_{SND}$

(c) $\hat{Q}(\hat{A})_{\hat{g}}$

Figure 4.9: Histograms of p-values for complex scenario 3 ($g(\boldsymbol{x}_i) = 5 + e_i$)



(a) $\hat{Q}(\hat{A})_{RS}$

(b) $\hat{Q}(\hat{A})_{SND}$

(c) $\hat{Q}(\hat{A})_{\hat{g}}$

Figure 4.10: Histograms of p-values for complex scenario 3 ($g(\boldsymbol{x}) = 0$)

CHAPTER 5

Discussion and future work

We proposed two methods which use randomized clinical trial data to identify subgroups of enhanced treatment effect. The first method was a penalized monotone single-index model, and could be used for subgroup identification in a variety of ways. For instance, one could consider using this model for the variable selection stage of an existing subgroup identification procedure, such as Virtual Twins [Foster et al., 2011]. That is, this model could be used as a replacement for the single regression tree in the second stage of the Virtual Twins procedure. One could also consider using a penalized monotone single-index model to estimate the treatment effect directly in the Average Value procedure, converting it to a one-stage procedure, and eliminating the need for the computationally expensive subgroup search. By penalizing the index parameter in this model, we are able to greatly reduce model complexity, which, combined with the monotonicity constraint on the function η, means the resulting subgroup will generally be relatively easy to understand. The second method is the Average Value procedure mentioned above, which can be viewed as a model-based alternative to the Virtual Twins procedure, and which was found to very effectively identify truly important covariates in our simulation study. Though this procedure is more computationally demanding than Virtual Twins, and was in some ways outperformed by Virtual Twins in our simulations, it has the advantage of identifying a single, contiguous region, which will often depend on fewer covariates than the subgroup identified by Virtual Twins, and may thus be somewhat easier to understand.

In addition to the subgroup identification procedures, we proposed a number of permutation-based methods for obtaining p-values for treatment-by-covariate interactions, and showed that they perform well compared to simple permutation tests, especially for more complex models and somewhat ad hoc test statistics. We also considered the use of these methods to obtain p-values for the enhancement metric discussed by Foster et al. [2011]. Having p-values along with enhancement metric estimates could help to further reduce the chances of falsely declaring an identified subgroup to be enhanced.

Overall, we feel the methods proposed in this dissertation represent a meaningful contribution to the field of subgroup analysis. We showed that, in many cases, very simple subgroups can accurately identify subjects who will show an enhanced response to treatment. Additionally, we considered a number of methods for evaluating the identified regions, such as permutation-based p-values for interaction tests. We believe such post-identification evaluation is important, as most subgroup identification procedures will nearly always identify a region, regardless of whether or not it is truly enhanced. Using the proposed methods, it may be possible to make more informed (and more confident) treatment decisions in the future using only a very limited amount of patient information. This lack of necessary information, along with the simple form (and resulting easy interpretability) of the subgroups identified by our methods, should lead to treatment assignment rules which will see a good deal of real-world use. Because of this, we believe the proposed methods may also encourage others working in the field of subgroup analysis to consider giving more weight to interpretability.

There are a variety of possible extensions to the methods considered in this dissertation. To further reduce the risk of false positives, we could consider the use of external information to define a subgroup. This could be done by developing an approach for weighting covariates which were pre-determined to be important, so that some are more likely than others to be chosen to define the "enhanced" region. Alternatively, we could consider the development of a method for pre-screening the covariates. Such a method could be used, say, after the first step of the Virtual Twins or Average Value procedures, so that only a subset of the covariates would be included as candidates for defining potential subgroups. This could be particularly helpful if we wish to consider other forms of data, such as genetic data, which could have thousands of covariates.

Thus far, we have considered the case where only two treatment options exist, but in many cases there may be several potentially good treatment options. Therefore, it would be useful to modify the proposed methods so that multiple treatment options could be considered. One simple way to do this would be to consider a different analysis for each unique treatment comparison, so that we would have several treatment effect estimates instead of one; however, this could dramatically increase the potential for false positive findings, and the results of such an analysis could be difficult to interpret. In some cases, one may know *a priori* which specific treatment comparisons are of interest, and could thus potentially reduce the number of unique treatment comparisons considered, thereby potentially reducing the risk for false findings and allowing for more interpretable results. For situations where one doesn't know which specific comparisons are of interest, it may also be interesting to consider the development of some sort of screening method which “weeds out” comparisons between similar treatments.

Another issue is the implementation of subroup identification procedures when there is missing data in the outcome, covariates or both. In such situations, multiple imputation is often employed, but combining inferences from multiple imputed data sets could be difficult when the output of a method is a subregion of the covariate space, rather than a numeric value, such as a parameter estimate or test statistic. One option would be to implement multiple imputation in the treatment effect estimation stage, but then use the observed data (with missingness) in the subgroup identification stage, but this is clearly not ideal. Thus, it would be very interesting to consider the development of a method for combining inferences from multiple imputed data sets when the object of interest is a subregion of the design space.

To this point, we have focused on using baseline information from RCT data to predict a patient’s response to treatment, but have not suggested how post-treatment information may be used. It would be very interesting to work on developing methods which use post-treatment information, in addition to baseline information, to better assign treatment to patients. This could be done within the framework of dynamic treatment regimes, in which treatment decisions at a given time are made using a patient’s history up to that time. A number of authors, including Murphy [2002], Lavori and Dawson [2004], Robins [2004], Laan and Petersen [2004], Murphy [2005], Laber et al. [2010], and Shortreed et al. [2011] have considered dynamic treatment regimes. Though very nice to work with, RCT data is often quite difficult to obtain, so the extension of the proposed methods to data from observational studies is also of interest.

APPENDIX

Appendix A

Asymptotics for Chapter 1

Statement of theorem

We establish the oracle properties for unconstrained adaptive lasso penalized single-index model estimates. Consider the following setup and regularity conditions of Hardle et al. [1993]. Assume the data $\{(\boldsymbol{x}_i, y_i) :, \ i = 1, \dots, n\}$ come from (2.1), where $\boldsymbol{\beta}_0$ is the true value of the index parameter, and the last $q < p$ elements of $\boldsymbol{\beta}_0$ are 0. Let $\mathcal{A} = \{j : \beta_{0,j} \neq 0\}$ and $\mathcal{A}_n^* = \{j : \hat{\beta}_j \neq 0\}$. Let $H \subseteq \mathbb{R}^p$ be a set chosen so denominators in formulas for kernel estimators are not too close to 0, where H is the union of a finite number of convex sets. Define $\boldsymbol{W}_0 = \begin{pmatrix} \boldsymbol{W}_{0(11)} & \boldsymbol{W}_{0(21)} \\ \boldsymbol{W}_{0(12)} & \boldsymbol{W}_{0(22)} \end{pmatrix}$ to be the $p \times p$ matrix:

$$\int\limits_H \big\{\boldsymbol{x} - E(\boldsymbol{X}_H | \boldsymbol{\beta}_0^T \boldsymbol{X}_H = \boldsymbol{\beta}_0^T \boldsymbol{x})\big\}\big\{\boldsymbol{x} - E(\boldsymbol{X}_H | \boldsymbol{\beta}_0^T \boldsymbol{X}_H = \boldsymbol{\beta}_0^T \boldsymbol{x})\big\}^T \times \eta'(\boldsymbol{\beta}_0^T \boldsymbol{x})^2 f(\boldsymbol{x}) d\boldsymbol{x},$$

where $\boldsymbol{X}$ is a random variable with the design density f, $\boldsymbol{W}_{0(11)}$ is $(p-q)\times(p-q)$ and $\boldsymbol{X}_H$ has the distribution of $\boldsymbol{X}$, conditional on $X \in H$. Given $\delta > 0$, let H^δ denote the set of all points in $\mathbb{R}^p$ distant no further than δ from H. Put $\mathscr{U} = \{\boldsymbol{\beta}_0^T \boldsymbol{x} : \boldsymbol{x} \in H^\delta\}$, and let ζ denote the density of $\boldsymbol{\beta}_0^T \boldsymbol{X}$. We define the following conditions for some $\delta > 0$:

1. f is bounded away from 0 on H^δ and has two bounded derivatives there;

2. η and ζ have two bounded, continuous derivatives on $\mathscr{U}$;

3. K is supported on the interval $(-1, 1)$ and is a symmetric probability density, with a bounded derivative;

4. $E(\epsilon_i|\boldsymbol{x}_i) = 0$, $E(\epsilon_i^2|\boldsymbol{x}_i) = \sigma^2(\boldsymbol{x}_i)$ for all i, where the function σ^2 is bounded and continuous and $sup_i E|\epsilon_i|^m = M_m < \infty$ for all m.

As noted by Hardle et al. [1993], the emphasis on two derivatives in (1) and (2) is motivated by the use of a second-order kernel, and the restriction in (1) that f be bounded away from 0 on H^δ ensures with high probability that the denominator in (2.3) is bounded away from zero for $t = \boldsymbol{\beta}^T \boldsymbol{x}$, where $\boldsymbol{x} \in H$ and $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}_0$. Let $\boldsymbol{B}$ denote the set of all unit p-vectors. Given $C > 0$, and $0 < C_1 < C_2 < \infty$, $\boldsymbol{B}_n = \{\boldsymbol{\beta} \in \boldsymbol{B} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq Cn^{-1/2}\}$, and $\mathscr{H}_n = \{h : C_1 n^{-1/5} \leq h \leq C_2 n^{-1/5}\}$. We assume that $\hat{\boldsymbol{\beta}} \in \boldsymbol{B}_n$ and $h \in \mathscr{H}_n$. This is likely to be true if we start with a $\sqrt{n}$-consistent estimator, such as that shown by Ichimura [1993] to exist for unpenalized, unconstrained SIM.

Theorem A.1 (Oracle Properties). *Let* $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ *be iid observations from model* (2.1) *such that conditions (1)-(4) hold. Suppose that* $\boldsymbol{W}_0$ *is positive-definite, and that* $\frac{\lambda_n}{\sqrt{n}} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$, *where* $\gamma \in (0, \frac{3}{5}]$. *Then the adaptive-LASSO penalized single-index model estimates must satisfy:*

1. *Consistency in variable selection:* $lim_n P(\mathcal{A}_n^* = \mathcal{A}) = 1$;

2. *Asymptotic normality:* $\sqrt{n} \times (\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) \rightarrow_d N(\mathbf{0}, \sigma^2 \boldsymbol{W}_{0(11)}^{-1})$.

Thus, the adaptive lasso penalized single-index model estimates perform as well as if the true nonzero elements of $\boldsymbol{\beta}_0$ were known.

Proofs

Theorem 1, part (b)

Let $\hat{S}(\boldsymbol{\beta}, h)$ be the non-penalized sum of squares. Under conditions (i)-(iv), we know from Hardle et al. [1993] that

$$\hat{S}(\boldsymbol{\beta}, h) = \tilde{S}(\boldsymbol{\beta}) + T(h) + R_1(\boldsymbol{\beta}, h) + R_2(h),$$

where $\sup_{\boldsymbol{\beta} \in \boldsymbol{B}_n, h \in \mathscr{H}_n} |R_1(\boldsymbol{\beta}, h)| = o_p(n^{1/5})$, and $\tilde{S}(\boldsymbol{\beta}) = n\{\boldsymbol{W}_0^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\}^T\{\boldsymbol{W}_0^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\} + R_4(\boldsymbol{\beta})$, where $\sup_{\boldsymbol{\beta} \in \boldsymbol{B}_n} |R_4(\boldsymbol{\beta})| = o_p(1)$. From this point on, we drop any terms not depending on $\boldsymbol{\beta}$, since $\boldsymbol{\beta}$ is our primary interest. Also, since R_4 is of smaller order than $\tilde{S}$, it is negligible. Thus, we can treat $\tilde{S}$ as our objective function. Considering now $\tilde{S}_2 = \tilde{S} + \lambda_n \sum \hat{w}_j |\beta_j|$, we can apply reasoning similar to Zou [2006]. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \frac{\boldsymbol{u}}{\sqrt{n}}$, where $\|\boldsymbol{u}\| \leq C$ and $\hat{w}_j = |\hat{\beta}_{init}|^{-\gamma}$. Then we have:

$$\tilde{S}_2(\boldsymbol{u}) = n\left\{\boldsymbol{W}_0^{\frac{1}{2}}\frac{\boldsymbol{u}}{\sqrt{n}} - \frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\right\}^T \left\{\boldsymbol{W}_0^{\frac{1}{2}}\frac{\boldsymbol{u}}{\sqrt{n}} - \frac{\sigma}{\sqrt{n}}\boldsymbol{Z}\right\} + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left|\beta_{0,j} + \frac{u_j}{\sqrt{n}}\right| + o_p(1).$$

Where $\boldsymbol{Z} \rightarrow_d \boldsymbol{T} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$. Let $\hat{\boldsymbol{u}} = \text{argmin } \tilde{S}_2(\boldsymbol{u})$; then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \frac{\hat{\boldsymbol{u}}}{\sqrt{n}}$, or $\hat{\boldsymbol{u}} = \sqrt{n} \times (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ Thus, we know $\tilde{S}_2(\boldsymbol{u}) - \tilde{S}_2(\boldsymbol{0}) = V^{(n)}(\boldsymbol{u})$, where

$$V^{(n)}(\boldsymbol{u}) =$$
$$\boldsymbol{u}^T \boldsymbol{W}_0^{\frac{1}{2}} \boldsymbol{W}_0^{\frac{1}{2}} \boldsymbol{u} - 2\sigma \boldsymbol{u}^T \boldsymbol{W}_0^{\frac{1}{2}} \boldsymbol{Z} + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^{p} \sqrt{n} \hat{w}_j \left(\left| \beta_{0,j} + \frac{u_j}{\sqrt{n}} \right| - |\beta_{0,j}| \right) + o_p(1).$$

The first term does not depend on n and we know that $\boldsymbol{Z} \rightarrow_d \boldsymbol{T} \sim N(\boldsymbol{0}, \boldsymbol{I}_p)$. The limiting behavior of the third term can be argued exactly as in Zou [2006]. In particular, if $\beta_{0,j} \neq 0$, then we know $\hat{w}_j \rightarrow_p |\beta_{0,j}|^{-\gamma}$ and $\sqrt{n}(|\beta_{0,j} + \frac{u_j}{\sqrt{n}}| - |\beta_{0,j}|) \rightarrow_p u_j \text{sign}(\beta_{0,j})$, so by Slutsky's theorem we know $\frac{\lambda_n}{\sqrt{n}} \hat{w}_j \sqrt{n}(|\beta_{0,j} + \frac{u_j}{\sqrt{n}}| - |\beta_{0,j}|) \rightarrow_p 0$, since $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$, as long as $u_j < \infty$. If $\beta_{0,j} = 0$, then $\sqrt{n}(|\beta_{0,j} + \frac{u_j}{\sqrt{n}}| - |\beta_{0,j}|) = |u_j|$ and $\frac{\lambda_n}{\sqrt{n}} \hat{w}_j = \frac{\lambda_n}{\sqrt{n}} n^{\gamma/2} (|\sqrt{n} \hat{\beta}_{init,j}|)^{-\gamma}$, where $\sqrt{n} \hat{\beta}_{init,j} = O_p(1)$. Therefore, using Slutsky's theorem again, we can see that $V^{(n)}(\boldsymbol{u}) \rightarrow_d V(\boldsymbol{u})$ for every $\boldsymbol{u}$, where

$$V(\boldsymbol{u}) = \begin{cases} \boldsymbol{u}_{\mathcal{A}}^T \boldsymbol{W}_{0(11)} \boldsymbol{u}_{\mathcal{A}} - 2\sigma \boldsymbol{u}_{\mathcal{A}}^T \boldsymbol{W}_{0(11)}^{\frac{1}{2}} \boldsymbol{T}_{\mathcal{A}} & \text{if } u_j = 0 \; \forall j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

Note that choosing u_j to be finite for $\beta_{0,j} \neq 0$ gives finite values of $V^{(n)}$, whereas if u_j is not finite for $\beta_{0,j} \neq 0$, $V^{(n)}$ is infinite, so the optimal $\boldsymbol{u}$ must be finite. We know that $V^{(n)}$ is convex (if $\boldsymbol{W}_0$ is positive-definite), and the unique minimum of V is $(\sigma \boldsymbol{W}_{0(11)}^{-\frac{1}{2}} \boldsymbol{T}_{\mathcal{A}}, \boldsymbol{0})$. Thus, following the epi-convergence argument used by Zou [2006], we have

$$\hat{\boldsymbol{u}}_{\mathcal{A}} \rightarrow_d \sigma \boldsymbol{W}_{0(11)}^{-\frac{1}{2}} \boldsymbol{T}_{\mathcal{A}} \text{ and } \hat{\boldsymbol{u}}_{\mathcal{A}^c} \rightarrow_d \boldsymbol{0}.$$

Since $\boldsymbol{T}_{\mathcal{A}} \sim N(\boldsymbol{0}, \boldsymbol{I})$, we know $\sqrt{n} \times (\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{0\mathcal{A}}) \rightarrow_d N(\boldsymbol{0}, \sigma^2 \boldsymbol{W}_{0(11)}^{-1})$, and we are done.

Theorem 1, part (a)

Again we can follow the general framework of Zou [2006]. We know that $\forall j \in \mathcal{A}$, the asymptotic normality results above indicate that $\hat{\beta}_j \to_p \beta_{0,j}$; thus meaning that $P(j \in \mathcal{A}_n^*) \to 1$. Therefore, we need only show that $\forall j' \notin \mathcal{A}$, $P(j' \in \mathcal{A}_n^*) \to 0$. Suppose that $j' \in \mathcal{A}_n^*$. By the KKT optimality conditions, we know that $2n\{\boldsymbol{W}_0^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - n^{-\frac{1}{2}}\sigma\boldsymbol{Z}\}^T\boldsymbol{W}_0^{\frac{1}{2}} = \lambda_n(\hat{w}_1\text{sign}(\hat{\beta}_1), \cdots, \hat{w}_p\text{sign}(\hat{\beta}_p))$, so $2n[\boldsymbol{W}_0^{\frac{1}{2}}]_{j'}^T\{\boldsymbol{W}_0^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - n^{-\frac{1}{2}}\sigma\boldsymbol{Z}\} = \lambda_n\hat{w}_{j'}\text{sign}(\hat{\beta}_{j'})$. We know generally that $|\text{sign}(t)| \to 1$ as $t \to 0$, and $\frac{\lambda_n}{\sqrt{n}}\hat{w}_{j'} = \frac{\lambda_n}{\sqrt{n}}n^{\gamma/2}(|\sqrt{n}\hat{\beta}_{init,j'}|)^{-\gamma} \to_p \infty$. However,

$$\frac{2}{\sqrt{n}}n[\boldsymbol{W}_0^{\frac{1}{2}}]_{j'}^T\{\boldsymbol{W}_0^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - n^{-\frac{1}{2}}\sigma\boldsymbol{Z}\} = 2[\boldsymbol{W}_0^{\frac{1}{2}}]_{j'}^T\boldsymbol{W}_0^{\frac{1}{2}}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - 2\sigma[\boldsymbol{W}_0^{\frac{1}{2}}]_{j'}^T\boldsymbol{Z},$$

where both terms on the right hand side converge in distribution to normals. Thus, we know

$$P(j' \in \mathcal{A}_n^*) \leq P\Big(2n[\boldsymbol{W}_0^{\frac{1}{2}}]_{j'}^T\{\boldsymbol{W}_0^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - n^{-\frac{1}{2}}\sigma\boldsymbol{Z}\} = \lambda_n\hat{w}_{j'}\text{sign}(\hat{\beta}_{j'})\Big) \to 0,$$

and we are done.

BIBLIOGRAPHY

BIBLIOGRAPHY

Susan F. Assmann, Stuart J. Pocock, Laura E. Enos, and Linda E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet*, 355 (9209):1064–1069, 2000.

Richard E. Barlow, David J. Bartholomew, John M. Bremner, and Hugh D. Brunk. *Statistical Inference Under Order Restrictions.* New York: John Wiley and Sons, 1972.

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

Jason Brinkley, Anastasios Tsiatis, and Kevin J. Anstrom. A generalized estimator of the attributable benefit of an optimal treatment regime. *Biometrics*, 66(2):512–522, 2010.

Sarah T. Brookes, Elise Whitley, Tim J. Peters, Paul A. Mulheran, Matthias Egger, and George Davey Smith. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. *Health technology assessment (Winchester, England)*, 5(33):1–56, 2001. ISSN 1366-5278.

Sarah T. Brookes, Elise Whitely, Matthias Egger, George Davey Smith, Paul A. Mulheran, and Tim J. Peters. Subgroup analyses in randomized trials: risks of subgroup-specific analyses; power and sample size for the interaction test. *Journal of Clinical Epidemiology*, 57(3):229–36, 2004. ISSN 0895-4356.

Petra Bůžková, Thomas Lumley, and Kenneth Rice. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of human genetics*, 75(1):36–45, January 2011. doi: 10.1111/j.1469-1809.2010.00572.x.

Tianxi Cai, Lu Tian, Peggy H. Wong, and L. J. Wei. Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12 (2):270–282, 2011.

Raymond J. Carroll, Jianqing Fan, Irene Gijbels, and Matt P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

Arindam Chatterjee and Soumendra N. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

Lu Cui, H. M. James Hung, Sue Jane Wang, and Yi Tsong. Issues related to subgroup analysis in clinical trials. *Journal of Biopharmaceutical Statistics*, 12(3):347–58, 2002. ISSN 1054-3406.

Eugene S. Edgington. *Randomization tests.* Marcel Dekker, Inc., New York, NY, USA, 1986.

Jianqing. Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–

1360, December 2001.

Jared C. Foster, Jeremy M.G. Taylor, and Stephen J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, pages 2867–2880, 2011.

Jerome Friedman and Robert Tibshirani. The monotone smoothing of scatterplots. *Technometrics*, 26(3):pp. 243–250, 1984.

Jerome Friedman, Trevor Hastie, Holger Hfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):pp. 302–332, 2007.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–141, 1991.

Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, 1999.

Wenjiang J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):pp. 397–416, 1998.

Phillip Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2000.

Lacey Gunter, Ji Zhu, and Susan Murphy. Variable selection for optimal decision making. In *Proceedings of the 11th conference on Artificial Intelligence in Medicine*, AIME '07, pages 149–154. Springer-Verlag, 2007.

Peter Hall and Li-Shan Huang. Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, 29(3):pp. 624–647, 2001.

Wolfgang Hardle, Peter Hall, and Hidehiko Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):pp. 157–178, 1993.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Xuming He and Peide Shi. Monotone b-spline smoothing. *Journal of the American Statistical Association*, 93(442):pp. 643–650, 1998.

Hidehiko Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1-2):71 – 120, 1993.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, Forthcoming, 2012.

Holly Janes, Margaret S. Pepe, Patrick M. Bossuyt, and William E. . Measuring the performance of markers for guiding treatment decisions. *Annals of Internal Medicine*, 154(4):253–259, 2011.

Stevo Julius, Shawna D. Nesbitt, Brent M. Egan, Michael A. Weber, Eric L. Michelson, Niko Kaciroti, Henry R. Black, Richard H. Grimm, Franz H. Messerli, Suzanne Oparil, and M. Anthony Schork. Feasibility of treating prehypertension with an angiotensin-receptor blocker. *New England Journal of Medicine*, 354(16):1685–1697, 2006.

Sidney Katz and C. Amechi Akpom. Index of ADL. *Medical Care*, 14(5):116–118, 1976.

Efang Kong and Yingcun Xia. Variable selection for the single index model. *Biometrika*, 94(1):217–229, 2007.

Mark van der Laan and Maya Petersen. History-adjusted marginal structural models and statically-optimal dynamic treatment regimes. U.C. Berkeley Division of Biostatistics Working Paper Series 1158, Berkeley Electronic Press, 2004. URL http://EconPapers.repec.org/RePEc:bep:ucbbio:1158.

Eric Laber, Min Qian, Dan J. Lizotte, and Susan A. Murphy. Statistical Inference in Dynamic Treatment Regimes. 2010. URL http://arxiv.org/abs/1006.5831.

Stephen W. Lagakos. The challenge of subgroup analyses–reporting without distorting. *New England Journal of Medicine*, 354(16):1667–9, 2006. ISSN 0028-4793.

Philip W. Lavori and Ree Dawson. Dynamic treatment regimes: practical design considerations. *Clinical Trials*, pages 9–20, 2004.

Ker-Chau Li and Naihua Duan. Regression analysis under link violation. *The Annals of Statistics*, 17(3):pp. 1009–1052, 1989.

Hua Liang, Xiang Liu, Runze Li, and Chih-Ling Tsai. Estimation and testing for partially linear single-index models. *Annals of Statistics*, 38(6):3811–3836, 2010.

Ilya Lipkovich, Alex Dmitrienko, Jonathan Denne, and Gregory Enas. Subgroup identification based on differential effect searcha recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30(21):2601–2621, 2011.

Enno Mammen. Estimating a smooth monotone regression function. *The Annals of Statistics*, 19(2):pp. 724–740, 1991.

Hari Mukerjee. Monotone nonparametric regression. *The Annals of Statistics*, 16(2): pp. 741–750, 1988.

Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65:331–366, 2002.

Susan A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005.

Abdissa Negassa, Antonio Ciampi, Michal Abrahamowicz, Stanley Shapiro, and Jean-François Boivin. Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15(3):231–239, 2005.

Richard Peto, Rory Collins, and Richard N. Gray. Large-scale randomized evidence: Large, simple trials and overviews of trials. *Journal of Clinical Epidemiology*, 48 (1):23–40, 1995. ISSN 0895-4356.

Stuart J. Pocock, Susan E. Assmann, Laura E. Enos, and Linda E. Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in Medicine*, 21(19):2917–2930, 2002. ISSN 0277-6715.

Richard F. Potthoff, Bercedis L. Peterson, and Stephen L. George. Detecting treatment-by-centre interaction in multi-centre clinical trials. *Statistics in Medicine*, 20(2):193–213, 2001.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180–1210, 2011.

James O. Ramsay. Monotone regression splines in action. *Statistical Science*, 3(4): pp. 425–441, 1988.

James M. Robins. Optimal structural nested models for optimal sequential decisions. In *In Proceedings of the Second Seattle Symposium on Biostatistics*. Springer, 2004.

Peter M. Rothwell. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet*, 365(9454):176–86, 2005.

Stephen J. Ruberg, Lei Chen, and Yanping Wang. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. *Clinical trials (London, England)*, 7(5):574–583, 2010. ISSN 1740-7753.

Susan M. Shortreed, Eric Laber, Daniel J. Lizotte, T. Scott Stroup, Joelle Pineau, and Susan A. Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine Learning*, 84(1-2):109–136, 2011.

Noah Simon and Robert Tibshirani. A permutation approach to testing interactions in many dimensions. 2012.

Xiao Song and Margaret Sullivan Pepe. Evaluating markers for selecting a patient's treatment. *Biometrics*, 60(4):pp. 874–883, 2004.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.*, 10:141–158, 2008.

Xiaogang Su, Tianni Zhou, Xin Yan, Juanjuan Fan, and Song Yang. Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4(1), 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996.

Rui Wang, Stephen W. Lagakos, James H. Ware, David J. Hunter, and Jeffrey M. Drazen. Statistics in medicine–reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–94, 2007.

Yingcun Xia and Wolfgang Karl Hrdle. Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97(5):1162–1184, 2006.

Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society Series B*, 64(3): 363–410, 2002.

Yan Yu and David Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):pp. 1042–1054, 2002.

Salim Yusuf, Janet Wittes, Jeffrey Probstfield, and Herman A. Tyroler. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA*, 266(1):93–98, 1991.

Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, and L. J. Wei. Effectively selecting a target population for a future comparative study. *Harvard University Biostatistics Working Paper Series*, Working Paper 134, 2011. URL `http://biostats.bepress.com/harvardbiostat/paper134`.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.

Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4):1733–1751, 2009.