A Simple Method for Detecting Interactions between a Treatment and a Large Number of Covariates

Lu Tian *

Ash A Alizadeh †

Andrew J Gentles ‡

and

Robert Tibshirani§

March 8, 2014

Abstract

We consider a setting in which we have a treatment and a potentially large number of covariates for a set of observations, and wish to model their relationship with an outcome of interest. We propose a simple method for modeling interactions between the treatment and covariates. The idea is to modify the covariate in a simple way, and then fit a standard model using the modified covariates and no main effects. We show that coupled with an efficiency augmentation procedure, this method produces clinically meaningful estimators in a variety of settings. It can be useful for practicing personalized medicine: determining from a large set of biomarkers the subset of

*Depts. of Health, Research & Policy, 94305, lutian@stanford.edu

†Dept. of Medicine, Stanford University. 94305, arasha@stanford.edu

‡Integrative Cancer Biology Program, Stanford University. andrewg@stanford.edu

§Depts. of Health, Research & Policy, and Statistics, Stanford University, tibs@stanford.edu

patients that can potentially benefit from a treatment. We apply the method to both simulated datasets and real trial data. The modified covariates idea can be used for other purposes, for example, large scale hypothesis testing for determining which of a set of covariates interact with a treatment variable.

1 Introduction

To develop strategies for personalized medicine, it is important to identify the treatment and covariate interactions in the setting of randomized clinical trials [Royston and Sauerbrei, 2008]. To confirm and quantify the treatment effect is often the primary objective of a randomized clinical trial. Although important, the final result (positive or negative) of a randomized trial is a conclusion with respect to the average treatment effect on the entire study population. For example, a treatment may not be different from the placebo in the overall study population, but it may still be better for a subset of patients. Identifying the treatment and covariate interactions may provide valuable information for determining this subgroup of patients.

In practice, there are two commonly used approaches to characterize the potential treatment and covariate interactions. First, a panel of simple patient subgroup analyses, where the treatment and control arms are compared in different patient subgroups defined a priori, such as male, female, diabetic and non-diabetic patients, may be performed following the main comparison. Such an exploratory approach is mainly focusing on simple interactions between treatment and a dichotomized covariate. However it often suffers from false positive findings due to multiple testings and can not find complicated treatment-covariates interaction.

In a more rigorous analytic approach, the treatment-covariates interactions can be examined in a multivariate regression analysis where the product of the binary treatment indicator

and a set of baseline covariates are included in the regression model. Recent breakthroughs in biotechnology make a vast amount of data available for exploring potential interaction effects with the treatment and assisting in the optimal treatment selection for individual patients. However, it is very difficult to detect the interactions between treatment and high dimensional covariates via the direct multivariate regression modeling. Appropriate variable selection methods such as Lasso are needed to reduce the number of covariates having interactions with the treatment. The presence of the main effect, which often have bigger effect on the outcome than the treatment interactions, further compounds the difficulty in dimension reduction since a subset of variables need to be selected for modeling the main effect as well.

Recently, Bonetti and Gelber [2004] formalized the subpopulation treatment effect pattern plot (STEPP) for characterizing interactions between the treatment and continuous covariates. Sauerbrei et al. [2007] proposed an efficient algorithm for multivariate model-building with flexible fractional polynomials interactions (MFPI) and compared the empirical performance of MFPI with STEPP. Su et al. [2008] proposed the classification and regression tree method to explore the covariates and treatment interactions in survival analysis. Tian and Tibshirani [2010] proposed an efficient algorithm to construct an index score, the sum of selected dichotomized covariates, to stratify the patient population according to the treatment effect. In a more recent work, Zhao et al. [2012] proposed a novel approach to directly estimate the optimal treatment selection rule via maximizing the expected clinical utility, which is equivalent to a weighted classification problem. There are also rich Bayesian literatures for flexible modeling nonlinear, nonadditive or interaction covariate effects [LeBlanc, 1995, Chipman et al., 1998, Gustafson, 2000, Chen et al., 2012]. However, most of these existing methods except for the one proposed by Zhao et al. [2012], are not designed to deal with high-dimensional covariates.

In this paper, we propose a simple approach to estimate the covariates and treatment

interactions without the need for modeling the main effects in analyzing data from a randomized clinical trial. The idea is simple, and in a sense, obvious. We simply code the treatment variable as ± 1 and then include the products of this variable with each covariate in an appropriate regression model.

Figure 1 gives a preview of the results of our method. The data consist of baseline covariates including various biomarker measurements and medical history for patients with stable coronary artery disease and normal or slightly reduced left ventricular function, who were randomized to either the angiotensin converting enzyme(ACE) inhibitor or placebo arm. Our proposed method constructs a numerical score using baseline information on a training set to reveal covariates-treatment interactions. The panels show the estimated survival curves for patients in a separate validation set, overall and stratified by the score. Although there is no significant survival difference between two arms in the overall comparison (hazard ratio=0.95, $p = 0.67$), we see that patients with low scores have better survival with the ACE inhibitor treatment than with the placebo (hazard ratio=0.74, $p = 0.06$). This type of information after appropriate validation could be very useful in clinical practice.

In section 2, we describe the methods for the continuous, binary as well as survival type of outcomes. We also establish a simple casual interpretation of the proposed method in several cases. In section 3, the finite sample performance of the proposed method has been investigated via extensive numerical studies. In section 4, we apply the proposed method to two real data examples. Finally, limitation and potential extensions of the method are discussed in section 5.

2 Method

In the following, we let $T = \pm 1$ be the binary treatment indicator and $Y^{(1)}$ and $Y^{(-1)}$ be the potential outcome if the patient received treatment $T = 1$ and -1, respectively. We

only observe $Y = Y^{(T)}$, T and $\mathbf{Z}$, a $q-$dimensional baseline covariate vector. Here we assume that the treatment is randomly assigned to a patient, i.e., T and $\mathbf{Z}$ are independent. The observed data consist of N independent and identically distributed copies of $(Y, T, \mathbf{Z})$, $\{(Y_i, T_i, \mathbf{Z}_i), i = 1, \cdots, N\}$. Furthermore, we let $\mathbf{W}(\cdot) : R^q \to R^p$ be a p dimensional functions of baseline covariates $\mathbf{Z}$ and always include an intercept. In practice, $\mathbf{W}(\cdot)$ may include spline basis functions and interactions selected by the users. We denote $\mathbf{W}(\mathbf{Z}_i)$ by $\mathbf{W}_i$ in the rest of the paper. Here the dimension of $\mathbf{W}_i$ could be large relative to the sample size N. For simplicity, we assume that $\text{Prob}(T = 1) = \text{Prob}(T = -1) = 1/2$.

2.1 Continuous Response Model

When Y is a continuous response, a simple multivariate linear regression model for characterizing the interactions between the treatment and covariates is

$$Y = \beta_0' \mathbf{W}(\mathbf{Z}) + \boldsymbol{\gamma}_0' \mathbf{W}(\mathbf{Z}) \cdot T/2 + \epsilon, \tag{1}$$

where ϵ is the mean zero random error. In this simple model, the interaction term $\boldsymbol{\gamma}_0' \mathbf{W}(Z) \cdot T$ models the heterogeneous treatment effect across the population and the linear combination of $\boldsymbol{\gamma}_0' \mathbf{W}(\mathbf{Z})$ can be used for identifying the subgroup of patients who may or may not benefit from the treatment. Note that since the vector $\mathbf{W}(\mathbf{Z})$ contains an intercept, the main effect for treatment is always included in the model. Specifically, under model (1), we have

$$\begin{aligned} \Delta(\mathbf{z}) &= \text{E}(Y^{(1)} - Y^{(-1)} | \mathbf{Z} = \mathbf{z}) \\ &= \text{E}(Y | T = 1, \mathbf{Z} = \mathbf{z}) - \text{E}(Y | T = -1, \mathbf{Z} = \mathbf{z}) \\ &= \boldsymbol{\gamma}_0' \mathbf{W}(\mathbf{z}), \end{aligned}$$

i.e., $\boldsymbol{\gamma}_0'\mathbf{W}(\mathbf{z})$ measures the causal treatment effect for patients with the baseline covariate $\mathbf{Z}$. With observed data, $\boldsymbol{\gamma}_0$ can be estimated along with β_0 via the ordinary least squares (OLS) method.

On the other hand, noting the relationship that

$$\mathrm{E}(2YT|\mathbf{Z}=\mathbf{z}) = \Delta(\mathbf{z}),$$

one may estimate $\boldsymbol{\gamma}_0$ by directly minimizing

$$N^{-1}\sum_{i=1}^{N}(2Y_iT_i - \boldsymbol{\gamma}'\mathbf{W}_i)^2. \tag{2}$$

We call this the *modified outcome* method, where $2YT$ can be viewed as the *modified outcome*, which has been first proposed in the Ph.D thesis of James Sinovitch, Harvard University.

Under the simple linear model (1), estimators from both methods are consistent for $\boldsymbol{\gamma}_0$, and the full least squares approach in general is more efficient than the modified outcome method. In practice, the simple multivariate linear regression model is often just a working model approximating the complicated underlying probabilistic relationship among the treatment, baseline covariates and outcome variables. It comes as a surprise that even when model (1) is mis-specified, the multivariate linear regression and modified outcome estimators still converge to the same deterministic limit $\boldsymbol{\gamma}^*$, which is a non-random vector determined by the joint distribution of $(Y^{(1)}, Y^{(-1)}, \mathbf{Z})$. Furthermore, the score $\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*$ is a sensible estimator for the interaction effect in the sense that it seeks the "best" function of $\mathbf{z}$ in a functional space $\mathcal{F}$ to approximate $\Delta(\mathbf{z})$ by solving the optimization problem:

$$\min_f \mathrm{E}\{\Delta(\mathbf{Z}) - f(\mathbf{Z})\}^2,$$

$$\text{subject to } f \in \mathcal{F} = \{\boldsymbol{\gamma}'\mathbf{W}(\mathbf{z})|\boldsymbol{\gamma} \in R^p\},$$

where the expectation is with respect to $\mathbf{Z}$.

2.2 The Modified Covariate Method

The modified outcomes estimator defined above is useful for the Gaussian case, but does not generalize easily to more complicated models. Hence we propose a new estimator which is equivalent to the modified outcomes approach in the Gaussian case and extends easily to other models. This is the main proposal of this paper.

We consider the simple working model

$$Y = \alpha_0 + \boldsymbol{\gamma}_0' \frac{\mathbf{W}(\mathbf{Z}) \cdot T}{2} + \epsilon, \tag{3}$$

where ϵ is the mean zero random error. Based on model (3), we propose the *modified covariate* estimator $\hat{\boldsymbol{\gamma}}$ as the minimizer of

$$\frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \boldsymbol{\gamma}' \frac{\mathbf{W}_i \cdot T_i}{2} \right)^2. \tag{4}$$

The fact that we can directly estimate $\boldsymbol{\gamma}_0$ in model (3) without considering the intercept α_0 is due to the orthogonality between $\mathbf{W}(\mathbf{Z}_i) \cdot T_i$ and the intercept, which again is the consequence of the randomization. That is, we simply multiply each component of $\mathbf{W}_i$ by one-half the treatment assignment indicator ($= \pm 1$) and perform a regular linear regression. Now since

$$\frac{1}{N} \sum_{i=1}^{N} \left\{ Y_i - \boldsymbol{\gamma}' \frac{\mathbf{W}_i \cdot T_i}{2} \right\}^2 = \frac{1}{4N} \sum_{i=1}^{N} \left\{ 2 Y_i T_i - \boldsymbol{\gamma}' \mathbf{W}_i \right\}^2,$$

the modified outcome and modified covariate estimates are identical and share the same causal interpretations. Operationally, we can omit the intercept and perform a simple linear regression with the modified covariates. In general, we proposed the following modified

covariate approach

1. Modify the covariate

$$Z_i \rightarrow \mathbf{W}_i = \mathbf{W}(\mathbf{Z}_i) \rightarrow \mathbf{W}_i^* = \mathbf{W}_i \cdot T_i/2$$

2. Perform appropriate regressions

$$Y \sim \boldsymbol{\gamma}_0' \mathbf{W}^* \tag{5}$$

based on the modified observations

$$(\mathbf{W}_i^*, Y_i) = \{(\mathbf{W}_i \cdot T_i)/2, Y_i\}, i = 1, 2, \ldots N. \tag{6}$$

3. $\hat{\boldsymbol{\gamma}}' \mathbf{W}(\mathbf{z})$ can be used to stratify patients for individualized treatment selection.

Figure 2 illustrates how the modified covariate method works for a single covariate Z in two treatment groups. The raw data are shown on the left and the data with modified covariate are shown on the right. The regression line computed in the right panel estimates the treatment-covariate interaction.

The advantage of this new approach is twofold: it avoids having to directly model the main effects and it has a causal interpretation for the resulting estimator regardless of the adequacy of the assumed working model (3). Furthermore, unlike the modified outcome method, it is straightforward to generalize the new approach to other types of outcome.

2.3 Binary Responses

When Y is a binary response, in the same spirit as the continuous outcome case, we propose to fit a multivariate logistic regression model with modified covariates $\mathbf{W}^* = \mathbf{W}(\mathbf{Z}) \cdot T/2$ generalized from (5):

$$\text{Prob}(Y = 1|\mathbf{Z}, T) = \frac{\exp(\boldsymbol{\gamma}_0' \mathbf{W}^*)}{1 + \exp(\boldsymbol{\gamma}_0' \mathbf{W}^*)}. \tag{7}$$

If model (7) is correctly specified, then

$$\begin{aligned} \Delta(\mathbf{z}) &= \text{Prob}(Y^{(1)} = 1|\mathbf{Z} = \mathbf{z}) - \text{Prob}(Y^{(-1)} = 1|\mathbf{Z} = \mathbf{z}) \\ &= \text{Prob}(Y = 1|T = 1, \mathbf{Z} = \mathbf{z}) - \text{Prob}(Y = 1|T = -1, \mathbf{Z} = \mathbf{z}) \\ &= \frac{\exp\{\boldsymbol{\gamma}_0' \mathbf{W}(\mathbf{z})/2\} - 1}{\exp\{\boldsymbol{\gamma}_0' \mathbf{W}(\mathbf{z})/2\} + 1}, \end{aligned}$$

and thus $\boldsymbol{\gamma}_0' \mathbf{W}(\mathbf{z})$ has an appropriate causal interpretation. However, even when model (7) is not correctly specified, we still can estimate $\boldsymbol{\gamma}_0$ by treating (7) as a working model. In general, the maximum likelihood estimator (MLE) of the working model converges to a deterministic limit $\boldsymbol{\gamma}^*$ and $\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2$ can be viewed as the solution to the following optimization problem

$$\max_f \text{E}\left\{Y f(\mathbf{Z}) T - \log(1 + e^{f(\mathbf{Z})T})\right\}$$

$$\text{subject to } f \in \mathcal{F} = \{\gamma' \mathbf{W}(\mathbf{z})/2 | \boldsymbol{\gamma} \in R^p\},$$

where the expectation is with respect to $(Y, T, \mathbf{Z})$. Therefore, if $\mathbf{W}(\mathbf{z})$ forms a "rich" set of basis functions, $\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2$ is an approximation to the minimizer of $\text{E}\left\{Y f(\mathbf{Z}) T - \log(1 + e^{f(\mathbf{Z})T})\right\}$. In the appendix 6.1, we show that the latter can be represented as

$$f^*(\mathbf{z}) = \log\left\{\frac{1 - \Delta(\mathbf{z})}{1 + \Delta(\mathbf{z})}\right\}$$

under very general assumptions. Therefore,

$$\hat{\Delta}(\mathbf{z}) = \frac{\exp\{\hat{\boldsymbol{\gamma}}'\mathbf{W}(\mathbf{z})/2\} - 1}{\exp\{\hat{\boldsymbol{\gamma}}'\mathbf{W}(\mathbf{z})/2\} + 1}$$

may serve as an estimate for the covariate-specific treatment effect and used to stratify the patient population, regardless of the validity of the working model assumptions.

As described above, the MLE from the working model (7) can always be used to construct a surrogate to the personalized treatment effect measured by the "risk difference"

$$\Delta(\mathbf{z}) = \mathrm{E}(Y^{(1)} - Y^{(-1)}|\mathbf{Z} = \mathbf{z}).$$

On the other hand, different measures for individualized treatment effects such as relative risk may also be of interest. For example, if we consider an alternative approach for fitting the logistic regression working model (7) by letting

$$\hat{\boldsymbol{\gamma}} = \mathrm{argmin}_{\boldsymbol{\gamma}} \sum_{i=1}^{N} \left\{ (1 - Y_i)\boldsymbol{\gamma}'\mathbf{W}^* + Y_i e^{-\boldsymbol{\gamma}'\mathbf{W}_i^*} \right\},$$

then $\hat{\boldsymbol{\gamma}}$ converges to a deterministic limit $\tilde{\boldsymbol{\gamma}}^*$ and $\exp\{\mathbf{W}(\mathbf{z})'\tilde{\boldsymbol{\gamma}}^*(\mathbf{z})/2\}$ can be viewed as an approximation to

$$\widetilde{\Delta}(\mathbf{z}) = \frac{\mathrm{Prob}(Y^{(1)} = 1|\mathbf{Z} = \mathbf{z})}{\mathrm{Prob}(Y^{(-1)} = 1|\mathbf{Z} = \mathbf{z})},$$

which measures the treatment effect using "relative risk" rather than "risk difference". This loss function is motivated by the fact that the logistic regression model (7) can be fitted by solving the estimating equation

$$N^{-1} \sum_{i=1}^{N} \left[\mathbf{W}_i^* \left\{ (1 - Y_i) - Y_i e^{-\boldsymbol{\gamma}'\mathbf{W}_i^*} \right\} \right] = 0,$$

which is the derivative of the proposed loss function. The detailed justification is given in

appendix 6.1.

2.4 Survival Responses

When the outcome variable is survival time, we often do not observe the exact outcome for every subject in a clinical study due to incomplete follow-up. In this case, we assume that the outcome Y is a pair of random variables $(X, \delta) = \{\tilde{X} \wedge C, I(\tilde{X} < C)\}$, where $\tilde{X}$ is the survival time of primary interest, C is the censoring time and δ is the censoring indicator.

Firstly, we propose to fit a Cox regression model with modified covariates

$$\lambda(t|\mathbf{Z}, T) = \lambda_0(t)e^{\boldsymbol{\gamma}'\mathbf{W}^*} \tag{8}$$

where $\lambda(t|\cdot)$ is the hazard function for survival time $\tilde{X}$ and $\lambda_0(\cdot)$ is a baseline hazard function free of $\mathbf{Z}$ and T. When model (8) is correctly specified,

$$\begin{aligned}
\Delta(\mathbf{z}) &= \frac{\mathrm{E}\{\Lambda_0(\tilde{X}^{(1)})|\mathbf{Z} = \mathbf{z}\}}{\mathrm{E}\{\Lambda_0(\tilde{X}^{(-1)})|\mathbf{Z} = \mathbf{z}\}} \\
&= \frac{\mathrm{E}\{\Lambda_0(\tilde{X})|T = 1, \mathbf{Z} = \mathbf{z}\}}{\mathrm{E}\{\Lambda_0(\tilde{X})|T = -1, \mathbf{Z} = \mathbf{z}\}} \\
&= \exp\{-\boldsymbol{\gamma}_0'\mathbf{W}(\mathbf{z})\}
\end{aligned}$$

and $\boldsymbol{\gamma}_0'\mathbf{W}(\mathbf{z})$ can be used to stratify the patient population according to $\Delta(\mathbf{z})$, where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is a monotone increasing function (the baseline cumulative hazard function). Under the proportional hazards assumption, the maximum partial likelihood estimator $\hat{\boldsymbol{\gamma}}$ is a consistent estimator for $\boldsymbol{\gamma}_0$ and semiparametric efficient. Moreover, even when model (8) is mis-specified, we still can "estimate" $\boldsymbol{\gamma}_0$ by maximizing the partial likelihood function. In general, the resulting estimator, $\hat{\boldsymbol{\gamma}}$, converges to a deterministic limit $\boldsymbol{\gamma}^*$, which is the root of a limiting score equation [Lin and Wei, 1989]. More generally, $\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2$ can be viewed

as the solution of the optimization problem

$$\max_f \mathrm{E} \int_0^\tau \left(f(\mathbf{Z})T - \log[\mathrm{E}\{e^{f(\mathbf{Z})T} I(X \geq u)\}] \right) dN(u)$$

$$\text{subject to } f \in \mathcal{F} = \{\boldsymbol{\gamma}' \mathbf{W}(\mathbf{z})/2 | \boldsymbol{\gamma} \in R^p\},$$

where $N(t) = I(\tilde{X} \leq t)\delta$, τ is a fixed time point such that $\mathrm{Prob}(X \geq \tau) > 0$, and the expectations are with respect to $(Y, T, \mathbf{Z})$. Therefore, $\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2$ can be viewed as an approximation to

$$f^*(\mathbf{z}) = \mathrm{argmax}_f \mathrm{E} \int_0^\tau \left(f(\mathbf{Z})T - \log[\mathrm{E}\{e^{f(\mathbf{Z})T} I(X \geq u)\}] \right) dN(u).$$

In appendix 6.1, we show that if the censoring time does not depend on $(\mathbf{Z}, T)$, then the minimizer f^* satisfies

$$\begin{aligned} & e^{f^*(\mathbf{z})} \mathrm{E}\{\Lambda^*(\tilde{X}^{(1)}) | \mathbf{Z} = \mathbf{z}\} - e^{-f^*(\mathbf{z})} \mathrm{E}\{\Lambda^*(\tilde{X}^{(-1)}) | \mathbf{Z} = \mathbf{z}\} \\ =& \mathrm{Prob}(\delta = 1 | T = 1, \mathbf{Z} = \mathbf{z}) - \mathrm{Prob}(\delta = 1 | T = -1, \mathbf{Z} = \mathbf{z}) \end{aligned}$$

for a monotone increasing function $\Lambda^*(u)$. Thus, when censoring rates are balanced between the two arms, e.g., the hazard rate is low and most of the censoring is due to administrative reasons,

$$f^*(\mathbf{z}) \approx -\frac{1}{2} \log \left[\frac{\mathrm{E}\{\Lambda^*(\tilde{X}^{(1)}) | \mathbf{Z} = \mathbf{z}\}}{\mathrm{E}\{\Lambda^*(\tilde{X}^{(-1)}) | \mathbf{Z} = \mathbf{z}\}} \right]$$

can be used to characterize the covariate-specific treatment effect and stratify the patient population even when the working model (8) is mis-specified.

2.5 Regularization for the High Dimensional Data

When the dimension of $\mathbf{W}^*$, p, is high, we can easily apply appropriate variable selection procedures based on the corresponding working model. For example, L_1 penalized (Lasso)

estimators proposed by Tibshirani [1996] can be directly applied to the modified data (6). In general, one may estimate $\boldsymbol{\gamma}$ by minimizing

$$\frac{1}{N}\sum_{i=1}^{N} l(Y_i, \boldsymbol{\gamma}'\mathbf{W}_i^*) + \lambda_N \|\boldsymbol{\gamma}\|_1, \tag{9}$$

where $\|\boldsymbol{\gamma}\|_1 = \sum_{j=1}^{p} |\gamma_j|$ and

$$l(Y_i, \boldsymbol{\gamma}'\mathbf{W}_i^*) = \begin{cases} \frac{1}{2}(Y_i - \boldsymbol{\gamma}'\mathbf{W}_i^*)^2 & \text{for continuous responses} \\ -\{Y_i \boldsymbol{\gamma}'\mathbf{W}_i^* - \log(1 + e^{\boldsymbol{\gamma}'\mathbf{W}_i^*})\} & \text{for binary responses} \\ -\left[\boldsymbol{\gamma}'\mathbf{W}_i^* - \log\{\sum_{j=1}^{N} e^{\boldsymbol{\gamma}'\mathbf{W}_j^*} I(X_j \geq X_i)\}\right]\delta_i & \text{for survival responses} \end{cases}.$$

The resulting lasso regularized estimator may share the appealing finite sample oracle properties [Van De Geer, 2008, Zhang and Huang, 2008, Van De Geer and Buhlmann, 2009, Negahban et al., 2012]. For example, for continuous outcomes, if we select the penalty parameter $\lambda_n = O(\{\log(p)/N\}^{1/2})$ and the covariates $\mathbf{W}(\mathbf{Z}_i)T_i/2, i = 1, \cdots, N$ satisfy the restricted eigenvalue condition, then

$$\text{Prob}\left(\|\hat{\boldsymbol{\gamma}}_{\lambda_N} - \boldsymbol{\gamma}^*\|_2^2 \leq c_3(\#\boldsymbol{\gamma}^*)\left\{\frac{\log(p)}{N}\right\}\right) \geq 1 - c_1 e^{-c_2 N \lambda_N^2},$$

where $\|\boldsymbol{\gamma}\|_2^2 = \sum_{j=1}^{p} \gamma_j^2$, $\hat{\boldsymbol{\gamma}}_{\lambda_N}$ is the corresponding lasso regularized estimator, $\#\boldsymbol{\gamma}^*$ is the number of non-zero components in the vector $\boldsymbol{\gamma}^*$, and $c_i, i = 1, 2, 3$ are positive constants depending on the joint distribution of $(Y, \mathbf{W}(\mathbf{Z}), T)$ [Negahban et al., 2012]. Therefore, although the dimension p may be big, $\boldsymbol{\gamma}^*$ can always be approximated by the lasso estimator with an approximation error proportional to only the number of its non-zero components, which is small if $\boldsymbol{\gamma}^*$ is sparse. Following the Corollary 3 of Negahban et al. [2012], similar results hold when $\boldsymbol{\gamma}^*$ is not exactly sparse but can be approximated by a sparse vector. One consequence is that if $\mathbf{W}(\cdot)$ is adequately rich such that $|\boldsymbol{\gamma}^*\mathbf{W}(\mathbf{z}) - f^*(\mathbf{z})|$ is small and $\boldsymbol{\gamma}^*$ is

either exact or approximately sparse, then $\hat{\gamma}'_{\lambda_N}\mathbf{W}(\mathbf{z})$ is a good approximation of $f^*(\mathbf{z})$ with a high probability, where $f^*(\mathbf{z})$ is a monotone transformation of the individualized treatment effect $\Delta(\mathbf{z})$. This property holds for any pair of finite N and p with a sufficiently small ratio $\log(p)/N$ and does not depend on the correct specification of the working model. A similar result for the expected utility in the patient population with the estimated optimal treatment assignment rule has been established in Qian and Murphy [2011]. This result can be extended to the binary case, with appropriate regularity condition on the covariates $\mathbf{W}(\mathbf{Z}_i)T_i/2$ to ensure that the negative log-likelihood function satisfies a form of restricted strong convexity [Negahban et al., 2012]. For the survival case, parallel results have been established for lasso-regularized maximum partial likelihood estimator under the assumption that the multiplicative Cox model is correctly specified [Kong and Nan, 2012, Huang et al., 2013]. The finite sample properties of the lasso estimator for mis-specified Cox model are still unknown and warrants further study.

It might be reasonable to suppose that the covariates interacting with the treatment will more likely be the ones exhibiting important main effects themselves. Therefore, one could also apply the adaptive Lasso procedure [Zou, 2006] with feature weights $\hat{w}_j$ proportional to the reciprocal of the univariate "association strength" between the outcome Y and the jth component of $\mathbf{W}(\mathbf{Z})$. Specifically, one may modify the penalty in (9) as

$$\lambda_N \sum_{j=1}^{p} \frac{|\gamma_j|}{\hat{w}_j}, \tag{10}$$

where $\hat{w}_j = |\hat{\theta}_i|^{-1}$ or $(|\hat{\theta}_{-1i}| + |\hat{\theta}_{1i}|)^{-1}$. Here $\hat{\theta}_{j1}$, $\hat{\theta}_{j(-1)}$ and $\hat{\theta}_j$, are the estimated regression coefficients of the jth component of $\mathbf{W}(\mathbf{Z})$ in appropriate univariate regression analysis with observations from the group $T = 1$, the group $T = -1$, and the combined group, respectively. Other regularization methods such as elastic net may also be used [Zou and Hastie, 2005].

Interestingly, one can treat the modified data (6) just as generic data and hence couple

it with other statistical learning techniques. For example, one can apply a classifier such as prediction analysis of microarrays (PAM) to the modified data for the purpose of finding subgroup of samples in which the treatment effect is large. We can also do large scale hypothesis testings on the modified data to determine which gene-treatment interactions have significant effects on the outcome.

2.6 Efficiency Augmentation

When the models (5, 7 and 8) with modified covariates are correctly specified, the MLE estimator for $\boldsymbol{\gamma}^*$ is the most efficient estimator asymptotically. However, when models are treated as working models subject to mis-specification, a more efficient estimator can be obtained for estimating the same $\boldsymbol{\gamma}^*$. To this end, note that in general $\hat{\boldsymbol{\gamma}}$ is defined as the minimizer of an objective function motivated from a working model:

$$\hat{\boldsymbol{\gamma}} = \text{argmin}_{\boldsymbol{\gamma}} \frac{1}{N} \sum_{i=1}^{N} l(Y_i, \boldsymbol{\gamma}' \mathbf{W}_i^*). \tag{11}$$

Since for any function $\mathbf{a}(\mathbf{z}) : R^q \to R^p$, $E\{T_i \mathbf{a}(\mathbf{Z}_i)\} = 0$ due to randomization, the minimizer of the augmented objective function

$$\frac{1}{N} \sum_{i=1}^{N} \{l(Y_i, \boldsymbol{\gamma}' \mathbf{W}_i^*) - T_i \mathbf{a}(\mathbf{Z}_i)' \boldsymbol{\gamma}\}$$

converges to the same limit as $\hat{\boldsymbol{\gamma}}$. Furthermore, by selecting an optimal augmentation term $\mathbf{a}(\cdot)$, the minimizer of the augmented objective function may have smaller variance than that of the minimizer of the original objective function. In appendix 6.2, we show that

$$\mathbf{a}_0(\mathbf{z}) = -\frac{1}{2}\mathbf{W}(\mathbf{z})\text{E}(Y|\mathbf{Z} = \mathbf{z}) \quad \text{and} \quad \mathbf{a}_0(\mathbf{z}) = -\frac{1}{2}\mathbf{W}(\mathbf{z})\{\text{E}(Y|\mathbf{Z} = \mathbf{z}) - 0.5\}$$

are optimal choices for continuous and binary responses, respectively. Therefore, we propose the following two-step procedures for estimating $\boldsymbol{\gamma}^*$:

1. Estimate the optimal $\mathbf{a}_0(\mathbf{z})$:

   (a) For continuous responses, fit the linear regression model $E(Y|\mathbf{Z}) = \xi'\mathbf{B}(\mathbf{Z})$ for the appropriate function $\mathbf{B}(\mathbf{Z})$ with OLS. Appropriate regularization will be used if the dimension of $\mathbf{B}(\mathbf{Z})$ is high. Let

$$\hat{\mathbf{a}}(\mathbf{z}) = -\frac{1}{2}\mathbf{W}(\mathbf{z}) \times \hat{\xi}'\mathbf{B}(\mathbf{z}).$$

   (b) For binary response, fit the logistic regression model $\text{logit}\{\text{Prob}(Y = 1|\mathbf{Z})\} = \xi'\mathbf{B}(\mathbf{Z})$ for the appropriate function $\mathbf{B}(\mathbf{Z})$ by maximizing the likelihood function. Let

$$\hat{\mathbf{a}}(\mathbf{z}) = -\frac{1}{2}\mathbf{W}(\mathbf{z}) \times \left\{ \frac{e^{\hat{\xi}'\mathbf{B}(\mathbf{z})}}{1 + e^{\hat{\xi}'\mathbf{B}(\mathbf{z})}} - \frac{1}{2} \right\}.$$

   Here $\mathbf{B}(\mathbf{z}) = \{B_1(\mathbf{z}), \cdots, B_S(\mathbf{z})\}$ and $\{B_s(\mathbf{z}) : R^q \to R^1, 1 \leq s \leq S\}$ is a set of basis functions selected by users for capturing the relationship between Y and $\mathbf{Z}$. It may include, for example, spline basis functions, interactions or other transformations of the original covariates. In practice, $\mathbf{B}(\mathbf{z})$ is not necessarily but can be the same as $\mathbf{W}(\mathbf{z})$.

2. Estimate $\boldsymbol{\gamma}^*$

   (a) For continuous responses, we minimize

$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{1}{2}(Y_i - \boldsymbol{\gamma}'\mathbf{W}_i^*)^2 - \boldsymbol{\gamma}'\hat{\mathbf{a}}(\mathbf{Z}_i)T_i \right\}.$$

   (b) For binary responses, we minimize

$$\frac{1}{N}\sum_{i=1}^{N} \left[-\{Y_i\boldsymbol{\gamma}'\mathbf{W}_i^* - \log(1 + e^{\boldsymbol{\gamma}'\mathbf{W}_i^*})\} - \boldsymbol{\gamma}'\hat{\mathbf{a}}(\mathbf{Z}_i)T_i \right].$$

For survival outcome, the log-partial likelihood function is not a simple sum of i.i.d terms. However, in appendix 6.2, we show that the optimal choice of $\mathbf{a}(\mathbf{z})$ is

$$\mathbf{a}_0(\mathbf{z}) = -\frac{1}{2}\left[\frac{1}{2}\mathbf{W}(\mathbf{z})\left\{G_1(\tau;\mathbf{z}) + G_2(\tau;\mathbf{z})\right\} - \int_0^\tau \mathbf{R}(u;\boldsymbol{\gamma}^*)\{G_1(du;\mathbf{z}) - G_2(du;\mathbf{z})\}\right],$$

where $G_1(u;\mathbf{z}) = E\{M(u)|\mathbf{Z}=\mathbf{z}, T=1\}$, $G_2(u;\mathbf{z}) = E\{M(u)|\mathbf{Z}=\mathbf{z}, T=-1\}$,

$$M(t,\mathbf{W}^*,\boldsymbol{\gamma}) = N(t) - \int_0^t \frac{I(X\geq u)e^{\boldsymbol{\gamma}'\mathbf{W}^*}d\mathrm{E}\{N(u)\}}{\mathrm{E}\{e^{\boldsymbol{\gamma}'\mathbf{W}^*}I(X\geq u)\}}$$

and

$$\mathbf{R}(u;\boldsymbol{\gamma}) = \frac{\mathrm{E}\{\mathbf{W}^* e^{\boldsymbol{\gamma}'\mathbf{W}^*}I(X\geq u)\}}{\mathrm{E}\{e^{\boldsymbol{\gamma}'\mathbf{W}^*}I(X\geq u)\}}.$$

Unfortunately, $\mathbf{a}_0(\mathbf{z})$ depends on the unknown parameter $\boldsymbol{\gamma}^*$. On the other hand, for the high-dimensional case, the interaction effect is usually small and it is not unreasonable to assume that $\boldsymbol{\gamma}^* \approx 0$. Furthermore, if the censoring patterns are similar in both arms, we have $G_1(u,\mathbf{z}) \approx G_2(u,\mathbf{z})$. Using these two approximations, we can simplify the optimal augmentation term as

$$\mathbf{a}_0(\mathbf{z}) = -\frac{1}{4}\mathbf{W}(\mathbf{z})\left\{G_1(\tau;\mathbf{z}) + G_2(\tau;\mathbf{z})\right\} = -\frac{1}{2}\mathbf{W}(\mathbf{z}) \times \mathrm{E}\{M(\tau)|\mathbf{Z}=\mathbf{z}\)$$

where

$$M(t) = N(t) - \int_0^t \frac{I(X\geq u)d\mathrm{E}\{N(u)\}}{\mathrm{E}\{I(X\geq u)\}}.$$

Therefore, we propose to employ the following approach to implement the efficiency augmentation procedure:

1. Calculate

$$\hat{M}_i(\tau) = N_i(\tau) - \int_0^\tau \frac{I(X_i \geq u) d\{\sum_{j=1}^N N_j(u)\}}{\sum_{j=1}^N I(X_j \geq u)}$$

for $i = 1, \cdots, N$ and fit the linear regression model $E(\hat{M}(t)|\mathbf{Z}) = \xi' \mathbf{B}(\mathbf{Z})$ for the appropriate function $\mathbf{B}(\mathbf{Z})$ with OLS and appropriate regularization if needed. Let

$$\hat{\mathbf{a}}(\mathbf{z}) = -\frac{1}{2}\mathbf{W}(\mathbf{z}) \times \hat{\xi}' \mathbf{B}(\mathbf{z}).$$

2. Estimate $\boldsymbol{\gamma}^*$ by minimizing

$$\frac{1}{N}\sum_{i=1}^N \left(- \left[\boldsymbol{\gamma}' \mathbf{W}_i^* - \log\{\sum_{j=1}^N e^{\boldsymbol{\gamma}' \mathbf{W}_i^*} I(X_j \geq X_i)\} \right] \Delta_i - \boldsymbol{\gamma}' \hat{\mathbf{a}}(\mathbf{Z}_i) T_i \right)$$

with appropriate penalization if needed.

Remark 1

When the response is continuous, the efficient augmentation estimator is the minimizer of

$$\begin{aligned} &\sum_{i=1}^N \left[\frac{1}{2} \left\{ Y_i - \frac{1}{2}\boldsymbol{\gamma}' \mathbf{W}(\mathbf{Z}_i) T_i / 2 \right\}^2 - \boldsymbol{\gamma}' \hat{\mathbf{a}}(\mathbf{Z}_i) T_i \right] \\ =&\sum_{i=1}^N \frac{1}{2} \left\{ Y_i - \hat{\xi}' \mathbf{B}(\mathbf{Z}_i) - \frac{1}{2}\boldsymbol{\gamma}' \mathbf{W}(\mathbf{Z}_i) T_i \right\}^2 + \text{constant}. \end{aligned}$$

This equivalence implies that this efficiency augmentation procedures is asymptotically equivalent to that based on a simple multivariate regression with main effect $\hat{\xi}' \mathbf{B}(\mathbf{Z})$ and interaction $\boldsymbol{\gamma}' \mathbf{W}(\mathbf{Z}) \cdot T$. This is not a surprise. As we pointed out in section 2.1, the choice of the main effect in the linear regression does not affect the asymptotical consistency of

estimating the interactions. On the other hand, a good choice of main effect model can help estimate the interaction, i.e., personalized treatment effect, more accurately.

Another consequence is that one may directly use the same standard algorithm to obtain the augmented estimator when the lasso penalty is used. For binary or survival responses, the augmented estimator under the lasso regularization can be obtained with slightly modified algorithm designed for the lasso optimization as well. The detailed algorithm is given in appendix 6.3.

Remarks 2

For nonlinear models such as logistic and Cox regressions, the augmentation method is NOT equivalent to the full regression approach including both the main effect and interaction terms. In those cases, different specifications of the main effects in the full regression model result in asymptotically different estimates for the interaction terms, which, unlike the proposed modified covariate estimators, in general can not be interpreted as surrogates for the personalized treatment effects.

Remark 3

With binary responses, the estimating equation targeting on approximating the relative risk is

$$N^{-1}\sum_{i=1}^{N}\mathbf{W}_i^*\{(1-Y_i)-Y_i e^{-\boldsymbol{\gamma}'\mathbf{W}_i^*}\}=0$$

and the optimal augmentation term $a_0(\mathbf{z})$ can be approximated by

$$-\frac{1}{2}\mathbf{W}(\mathbf{z})\left\{\mathrm{E}(Y|\mathbf{Z}=\mathbf{z})-\frac{1}{2}\right\}$$

when $\boldsymbol{\gamma}^* \approx 0$. The efficiency augmentation algorithm can be carried out accordingly.

Remark 4

The similar technique can also be used for improving other estimators such as that proposed by Zhao et al. [2012], where the surrogate objective function for the weighted misclassification error can be written in the form of (2.6) as well. The optimal function $\mathbf{a}_0(\mathbf{z})$ needs to be derived case by case.

3 Numerical Studies

In this section, we perform extensive numerical studies to investigate the finite sample performance of the proposed method in various settings: the treatment may or may not have the marginal main effect; the personalized treatment effect may depend on complicated functions of covariates such as interactions among covariates; the regression model for detecting the interactions may or may not be correctly specified. Due to the limitation of the space, we only present simulation results from the selected representative cases. The results for other scenarios are similar to those presented.

3.1 Continuous Responses

For continuous responses, we generated N independent Gaussian samples from the regression model

$$Y = (\beta_0 + \sum_{j=1}^{p} \beta_j Z_j)^2 + (\gamma_0 + \sum_{j=1}^{p} \gamma_j Z_j + \sum_{1 \le i < j \le p} \alpha_{ij} Z_i Z_j) T + \sigma_0 \cdot \epsilon, \tag{12}$$

where the covariates $(Z_1, \cdots, Z_p)$ follow a mean zero multivariate normal distribution with a compound symmetric variance-covariance matrix, $(1-\rho)\mathbf{I}_p + \rho\mathbf{1}'\mathbf{1}$, and $\epsilon \sim N(0,1)$. We let $(\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \cdots, \gamma_p) = (0.4, 0.8, -0.8, 0.8, -0.8, 0, \cdots, 0)$, $\alpha_{ij} = 0.8I(i = 1, j = 2)$, $\sigma_0 = \sqrt{2}$, $N = 100$, and $p = 50$ and 1000 representing high and low dimensional cases, respectively. The treatment T was generated as ± 1 with equal probability at random. We consider four sets of simulations:

1. $\beta_0 = (\sqrt{6})^{-1}, \beta_j = (2\sqrt{6})^{-1}, j = 3, 4, \cdots, 10$ and $\rho = 0$;
2. $\beta_0 = (\sqrt{6})^{-1}, \beta_j = (2\sqrt{6})^{-1}, j = 3, 4, \cdots, 10$ and $\rho = 1/3$;
3. $\beta_0 = (\sqrt{3})^{-1}, \beta_j = (2\sqrt{3})^{-1}, j = 3, 4, \cdots, 10$ and $\rho = 0$;
4. $\beta_0 = (\sqrt{3})^{-1}, \beta_j = (2\sqrt{3})^{-1}, j = 3, 4, \cdots, 10$ and $\rho = 1/3$.

Settings 1 and 2 represented cases with relatively small main effects, where the variances in response contributable to the main effect, interaction and random error were about 37.5%, 37.5% and 25%, respectively, when the covariates were correlated. Settings 3 and 4 represented cases with relatively big main effects, where the variances in response contributable to the main effect, interaction and random error were about 75%, 15% and 10%, respectively, when the covariates were correlated. For each of the simulated data set, we implemented three methods:

- *full regression:* Fit a multivariate linear regression with complete main effects and covariate/treatment interaction terms, i.e., the dimension of the covariate matrix was $2(p+1)$. The Lasso was used to select the variables.
- *new:* Fit a multivariate linear regression with the modified covariate $\mathbf{W}^* = (1, \mathbf{Z})' \cdot T/2$. The dimension of the covariate matrix was $p+1$. Again, the Lasso was used for selecting variables.

- *new/augmented:* The proposed method with efficiency augmentation, where $\mathrm{E}(Y|\mathbf{Z})$ was estimated with lasso-regularized ordinary least squared method and $\mathbf{B}(\mathbf{z}) = \mathbf{z}$.

The selection of $\mathbf{W}(\mathbf{z}) = \mathbf{B}(\mathbf{z}) = \mathbf{z}$ mimicked the realistic situation where the knowledge of the true functional forms for interaction and main effects were often lacking. For all three methods, we selected the Lasso penalty parameter via 20-fold cross-validation. To evaluate the performance of the resulting score measuring the individualized treatment effect, we estimated the Spearman's rank correlation coefficient between the estimated score and the "true" treatment effect

$$\Delta(\mathbf{Z}) = \mathrm{E}(Y^{(1)} - Y^{(-1)}|\mathbf{Z}) = 1.6 \times (0.5 + Z_1 - Z_2 + Z_3 - Z_4 + Z_1 Z_2)$$

in an independently generated set with a sample size of 10000. Based on 500 sets of simulations, we plotted the boxplots of the rank correlation coefficients between the estimated scores $\hat{\boldsymbol{\gamma}}'\mathbf{Z}$ and $\Delta(\mathbf{Z})$ under simulation settings (1), (2), (3) and (4) in the top left, top right, bottom left and bottom right panels of Figure 3, respectively. For the first three settings, the performance of the proposed method was better than that of the full regression approach. The superiority of the new method was fairly obvious especially when $p = 1000$. For example, in the setting 3, where all the covariates were independent and the main effect was relatively big, the median correlation coefficients were 0.14, 0.51 and 0.51 for the full regression, new and new/augmented methods, respectively. In the setting 4, the most challenging setting to estimate the individualized treatment effect, the median correlation coefficients were zero for all the methods. On the other hand, the proportions of obtaining a positive correlation coefficient are 41%, 48% and 46% for the full regression, new and new/augmented methods, respectively, while the corresponding proportions of obtaining a negative correlation coefficient, which corresponded to a false detection, were 34%, 20% and 17%, respectively. Therefore, the proposed method with and without augmentation was still

superior to the conventional counterpart. Furthermore, the new method was also superior in terms of selecting the right covariates interacting with the treatment. For example, in the setting 3 with $p = 1000$, while the proposed method (with or without augmentation) on average selected 8 covariates with 2 true positives, i.e. covariates from $Z_i, 1 \leq i \leq 4$, the full regression method only selected 5 covariates with one true positive. If p reduced to 50, then on average, the proposed method (with or without augmentation) selected 9.5 covariates with 4 true positives and the full regression method selected 11.5 covariates with 3 true positive.

3.2 Binary Responses

For binary responses, we used the same simulation design as that for the continuous response. Specifically, we generated N independent binary samples from the regression model

$$Y = I\left((\beta_0 + \sum_{j=1}^{p} \beta_j Z_j)^2 + (\gamma_0 + \sum_{j=1}^{p} \gamma_j Z_j + \sum_{1 \leq i < j \leq p} \alpha_{ij} Z_i Z_j)T + \sigma_0 \cdot \epsilon \geq 0\right), \quad (13)$$

where all the model parameters were the same as those in the case of continuous responses. Noting that the logistic regression model was mis-specified under the chosen simulation design. We also considered the same four settings with different combinations of β_j and ρ. For each of the simulated data set, we implemented three methods:

1. *full regression:* Fit a lasso regularized multivariate logistic regression with complete main effects and covariate/treatment interaction terms.

2. *new:* Fit a lasso regularized multivariate logistic regression (without intercept) with the modified covariate $\mathbf{W}^* = (1, \mathbf{Z})' \cdot T/2$.

3. *new/augmented:* The proposed method with efficiency augmentation, where $\mathrm{E}(Y|\mathbf{Z})$ was estimated with Lasso-penalized logistic regression.

To evaluate the performance of the resulting score measuring the individualized treatment effect, we estimated the Spearman's rank correlation coefficient between the estimated score and the "true" treatment effect

$$\Delta(\mathbf{Z}) = \mathrm{E}(Y^{(1)} - Y^{(-1)}|\mathbf{Z})$$

Although the scores measuring the interaction from the first and second/third methods were different even when the sample size goes to infinity, the rank correlation coefficients put them on the same footing in comparing performances.

In the top left, top right, bottom left and bottom right panels of Figure 4, we plotted the boxplots of the correlation coefficients between the estimated scores $\hat{\boldsymbol{\gamma}}'\mathbf{Z}$ and $\Delta(\mathbf{Z})$ under simulation settings (1), (2), (3) and (4), respectively. The patterns were similar to that for the continuous response. The "new/augmented method" performed the best or close to the best in all four settings. For example, in the setting 3, the median correlation coefficients were 0.02, 0.15 and 0.17 for the full regression, new and new/augmented methods, respectively, when $p = 1000$. In addition, while the proposed methods on average selected 15 covariates with 2 true positives, the full regression method only selected 2 covariates with 0.3 true positive in the same setting.

3.3 Survival Responses

For survival responses, we used the same simulation design as that for the continuous and binary responses. Specifically, we generated N independent survival time from the regression model

$$\tilde{X} = \exp\left\{(\beta_0 + \sum_{j=1}^{p}\beta_j Z_j)^2 + (\gamma_0 + \sum_{j=1}^{p}\gamma_j Z_j + \sum_{1\le i<j\le p}\alpha_{ij}Z_iZ_j)T + \sigma_0\cdot\epsilon\right\}, \quad (14)$$

where all the model parameters were the same as in the previous subsections. The censoring time was generated from the uniform distribution $U(0, \xi_0)$, where ξ_0 was selected to induce a censoring rate of 25%. For each simulated data set, we implemented the following three methods:

1. *full regression:* Fit a lasso regularized multivariate Cox regression with full main effect and covariate/treatment interaction terms, i.e., the dimension of the covariate matrix was $2p + 1$.

2. *new:* Fit a lasso regularized multivariate Cox regression with modified covariates.

3. *new/augmented:* The proposed method with efficiency augmentation. To model the $\mathrm{E}\{M(\tau)|\mathbf{Z}\}$, we used linear regression with the lasso regularization method.

To evaluate the performance of the resulting score measuring the individualized treatment effect, we estimated the Spearman's rank correlation coefficient between the estimated score and the "true" treatment effect based on survival probability at $t_0 = 20$

$$\Delta(\mathbf{Z}) = \mathrm{Prob}(\tilde{X}^{(1)} \geq t_0|\mathbf{Z}) - \mathrm{Prob}(\tilde{X}^{(-1)} \geq t_0|\mathbf{Z})$$

In the top left, top right, bottom left and bottom right panels of Figure 5, we plotted the boxplots of the correlation coefficients between the estimated scores $\hat{\boldsymbol{\gamma}}'\mathbf{Z}$ and $\Delta(\mathbf{Z})$ under simulation settings, (1), (2), (3) and (4), respectively. The patterns were similar to those for the continuous and binary responses and confirmed our findings that the "efficiency-augmented method" performed the best among the three methods. For example, when $p = 1000$, the median correlation coefficients were 0.00, 0.41 and 0.41 in the setting 3 for the full regression, new and new/augmented methods, respectively. In addition, on average the proposed method selected 11 covariates with 2.2 true positives and the full regression method only selected 2 covariates with 0.5 true positive in the same setting.

4 Examples

In this section, we applied the proposed method to analyze two real data examples. In the first example, we considered a recent clinical trial "Preventive of Events with Angiotensin Converting Enzyme Inhibition" (PEACE) to study if the ACE inhibitors are effective for lowering cardiovascular risk for patients with stable coronary artery disease and normal or slightly reduced left ventricular function [Braunwald et al., 2004]. In this study, 8290 patients were randomly assigned to treatment and control arms with 633 deaths occurred by the end of the study. The estimated hazard ratio is 0.95 with an insignificant p-value of 0.13. However, in a secondary analysis, Solomon et al. [2006] reported that ACE inhibitors might significantly reduce the mortality for patients whose kidney function was abnormal at the baseline. Although this result needs to be interpreted cautiously, it suggests the possibility of existence of a subgroup of patients who may benefit from the treatment. For this example, we considered the survival time as the primary endpoint and the objective was to use seven baseline covariates: age, gender, left ventricular ejection fraction (LVEF), renal function measured by EGFR, hypertension, diabetic status and history of myocardial infarction to build a scoring system capturing the individualized treatment effect. Those covariates were selected for their known association with the cardiovascular risk. The continuous covariates, age, LVEF and EGFR are log-transformed. In addition to the seven covariates, we also included all the two-way interactions among them in the model. In summary, $\mathbf{Z}$ was a 7 dimensional vector and $\mathbf{W}(\mathbf{Z})$ was a $7 + 7 \times 6/2 = 28$ dimensional vector. We included all patients with complete information on these seven covariates. The final data set consisted of 3947 patients in the treatment arm and 3918 patients in the placebo arm. The outcome of interest was the survival time and there were 292 and 315 deaths in the treatment and placebo arms, respectively. The estimated survival curves by arms were plotted in figure 6. The goal of the analysis was to construct a score using baseline covariates to identify

subgroups of patients who may or may not be benefited from the ACE inhibition treatment. To this end, we selected the first 2000 patients in the treatment arm and placebo arm to form the training set and reserved the rest 3865 patients as an independent validation set. In selecting the training and validation sets, we used the original order of the observations in the data set without additional sorting to ensure the objectivity.

First we maximized the partial likelihood function with modified covariates to construct a score aiming for capturing the individualized treatment effect. The resulting score was a linear combination of selected covariates and their two-way interactions. Here, a low score favored ACE inhibition treatment. We then applied the score to classify the patients in the validation set into the high and low score groups depending on whether the patient's score was greater than the median level. In the high score group, the survival time in the ACE inhibition arm was slightly shorter than that in the placebo arm with an estimated hazard ratio of 1.27 for ACE inhibitor versus placebo ($p = 0.163$). In the low score group, the survival time in the ACE inhibition arm was longer than that in the placebo arm with an estimated hazard ratio of 0.74 ($p = 0.061$). The estimated survival functions of both treatment arms were plotted in the upper panels of Figure 7. The interaction between the constructed score and treatment was statistically significant in the multivariate Cox regression based on the validation set ($p = 0.022$).

Furthermore, we implemented the efficiency augmentation method and obtained a new score. Again, we classified the patients in the validation set into the high and low score groups based on the constructed gene score. The results were very similar and the interaction between the constructed score and treatment was also statistically significant ($p = 0.025$).

For comparison purposes, we also fitted a multivariate Cox regression model with treatment, $\mathbf{W}(\mathbf{Z})$, and their interactions as the covariates. In this model, we had 57 different covariates. The resulting score failed to stratify the population according to the treatment effect in the validation set. The results were shown in the lower panel of Figure 7. The

interaction between the constructed score and treatment was not statistically significant ($p = 0.980$).

To further objectively examine the performance of the proposal in this data set, we randomly split the data into a training set of 2000 patients and a validation set of 5865 patients. We then estimated the score measuring individualized treatment effect in the training set with the modified covariates and full regression approaches. The lasso method coupled with BIC criterion was used in the full regression approach due to the large number of covariates relative to the number of deaths in the training set. Patients in the validation set were then stratified into the high and low score groups. We calculated the hazard ratios of the ACE inhibition arm versus the placebo arm in high and low score groups, respectively. In Figure 8, we plotted the boxplots of the hazard ratios in the high and low risk groups of the validation set based on 500 random splits. The results indicated that the proposed method tended to perform better than the commonly used full regression method in separating patients according to the treatment effect measured by the hazard ratio, which was consistent with our previous findings from the simulation studies. Furthermore, the empirical probability of obtaining an interaction significant at the one-sided 0.05 level in the validation set was 27.0% for the new method and 13.6% for the full regression method. This observation supported that our previous significant findings for the detected interaction was likely not due to the random chance.

It has been known that the breast cancer can be classified into different subtypes using gene expression profile and the effective treatment may be different for different subtypes of the disease [Loi et al., 2007]. In the second example, we applied the proposed method to study the potential interactions between gene expression levels and Tamoxifen treatment in breast cancer patients. The data set consisted of 414 patients in the cohort GSE6532 collected by Loi et al. [2007] for the purpose of characterizing ER-positive subtypes with gene expression profiles. The data set can be downloaded from the web-

site *www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6532*. Excluding patients with incomplete information, there were 268 and 125 patients receiving Tamoxifen and alternative treatments, respectively. In addition to the routine demographic information, we had $44,928$ gene expression measurements for each of the 393 patients. The outcome of the primary interest here was the distant metastasis free survival (MFS) time, which could be right censored due to incomplete follow-up. The MFS times were not statistically different between the two groups with a two-sided p value of 0.59 (Figure 9). The goal of the analysis was to construct a gene score using gene expression levels to identify a subgroup of patients who may benefit from the Tamoxifen treatment. We selected the first 90 patients from each group to form the training set and reserved the rest 213 patients as the independent validation set.

We identified 5,000 genes with the highest empirical variances and then constructed a gene score by fitting the Lasso penalized Cox regression model with modified covariates in the training set. The Lasso penalty parameter was selected via cross-validation. The resulting gene score was a linear combination of expression levels of seven genes. We applied the gene score to classify the patients in the validation set into the high and low score groups according to the median. In the high score group, the MFS time in the Tamoxifen group was shorter than that in the alternative group with an estimated hazard ratio of 3.52 ($p = 0.064$). In the low score group, the MFS time in the Tamoxifen group was longer than that in the alternative group with an estimated hazard ratio of 0.694 ($p = 0.421$). The estimated survival curves for both groups were plotted in the upper panels of Figure 10. The interaction between constructed score and treatment was statistically significant in the validation set ($p = 0.004$).

We implemented the efficiency augmentation method to obtain a new gene score, which was based on the expression level of eight genes. Again, we classified the patients in the validation set into the high and low score groups using the constructed score. The results were very similar to that from the gene score constructed without augmentation.

When we fitted a multivariate Cox regression model with treatment, the gene expression levels, and all treatment-gene interactions as the covariates, only one gene interacting with the treatment was selected by lasso. However, the interaction could not be reproduced in the validation set (lower panel of Figure 10). Furthermore, the computational speed was substantially slower due to the high-dimensional covariates matrix in this case.

The second example was chosen for demonstrating the potential use of the proposed method in the high-dimensional setting. An important limitation of this example is that the treatment was not randomly assigned to patients as in a standard randomized clinical trial and the gene expression levels were measured at the study baseline, which may be different from treatment initiation time. Therefore, the results need to be interpreted with caution and further verification based on data from a randomized clinical trial is desired.

5 Discussion

In this paper, we have proposed a simple method to explore the potential interactions between treatment and a set of high dimensional covariates. The general idea is to use $\mathbf{W}(\mathbf{Z}) \cdot T/2$ as new covariates in a regression model to predict the outcome. This provides a convenient approach for constructing a proper objective function whose optimizer can be used to estimate the individualized treatment effect as a function of given covariates. Once the objective function is constructed, one may employ one's favorite algorithms such as lasso and boosting to optimize the empirical version of the objective function and the resulting estimator can be used to stratify the patient population according to the treatment benefit. Therefore, the proposed method can be used in a much broader way than those already described in the paper. For example, after creating the modified covariates $\mathbf{W}(\mathbf{Z}) \cdot T/2$, data mining techniques such as nearest shrunken centroid classification, PAM and support vector machines can also be used to link the new covariates with the outcomes [Friedman, 1991,

Tibshirani et al., 2003, Hastie and Zhu, 2006]. Most dimension reduction methods in the literature can be readily adapted to handle the potentially high dimensional covariates. For univariate analysis, we also may perform large scale hypothesis testing on the modified data, to identify a list of covariates having interaction with the treatment; one could for example directly use the Significance Analysis of Microarrays (SAM) method [Gilbert et al., 2002] for this purpose. Extensions in these directions are promising and warrant further research.

As a limitation, the proposed method is primarily designed for analyzing data from randomized clinical trials. When applied to an observational study, where the covariates and treatment assignment are correlated, the constructed score may lose its causal interpretation. On the other hand, if a reasonable propensity score model is available, then we can still implement the modified covariate approach on matched or reweighted data so that the resulted score may retain the appropriate causal interpretations [Rosenbaum and Rubin, 1983].

Lastly, we want to emphasize that although the proposed method aims to estimate the individualized treatment effect with casual interpretation, the method is not immune to the common problems encountered in high dimensional data analysis such as multiple testing, false discovery, over-fitting et al., as the numerical studies demonstrated. To reinforce this point, we have repeated the simulation studies in Section 3 after removing the covariate/treatment interaction in generating the observed data and have found that there is non-negligible probability of false discoveries albeit the employment of cross-validation. In a typical example, where the covariates are independent and $p = 50$, the probabilities of detecting a non-existent interactions are as high as 52%, 44% and 40% for the full regression, new and new/augmented methods, respectively. Therefore the proposed method is just an exploratory tool and it is crucial to withhold an independent validation set, which can be used to verify the detected interaction. In the validation set, one may make valid statistical inference on various parameters of interest. For example, one may estimate and test the treatment effect in subgroup of patients selected from the detected covaraite/treatment

interactions and the results will have causal interpretation if the validation set is from a randomized clinical trial.

References

M. Bonetti and R. Gelber. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*, 5(3):465–81, 2004.

E Braunwald, M Domanski, S Fowler, N Geller, B Gersh, J Hsia, M Pfeffer, M Rice, Y Rosenberg, J Rouleau, and PEACE trial investigators. Angiotension-coverting-enzyme inhibition in stable coronary artery disease. *New England Journal of Medicine*, 351:2058–2068, 2004.

W. Chen, D. Ghosh, T. Raghunathan, M. Norkin, D. Sargent, and G. Bepler. On bayesian methods of exploring qualitative interactions for targeted treatment. *Statistics in Medicine*, 31, 2012.

H. Chipman, E. George, and R. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93, 1998.

J. Friedman. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 19(1):1–141, 1991.

C. Gilbert, B. Narasimhan, R. Tibshirani, and V. Tusher. Significance analysis of microarrays (sam) software. 2002. Available: http://www-stat.stanford.edu/~tibs/SAM/ via the Internet. Accessed 2003 July 16.

P. Gustafson. Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association*, 95(451):795–806, 2000.

T. Hastie and J. Zhu. Discussion of "support vector machines with applications" by Javier Moguerza and Alberto Munoz. *Statistical Science*, 21(3):352–357, 2006.

J Huang, T Sun, Z Ying, and C Zhang. Oracle inequalities for the lasso in the cox model. *Annals of Statistics*, 41:1055–1692, 2013.

S Kong and B Nan. Non-asymptotic oracle inequalities for the high-dimensional cox regression via lasso. *Tech Report*, page 1204.1992, 2012.

M. LeBlanc. An adaptive expansion method for regression. *Statistical Sinica*, 5, 1995.

D. Lin and LJ. Wei. Robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(107):1074–1078, 1989.

S. Loi, B. Haibe-kains, C. Desmedt, and et al. Definition of clinically distinct molecular subtypes in estrogen rceptor-positive breast carcinomas through genomic grade. *Journal of Clinical Oncology*, 25(10):1239–1246, 2007.

S. Negahban, P. Ravikumar, M. Wainwright, and B. Yu. A unified framework for high dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27:1214/12–STS400, 2012.

M Qian and S Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39:1180–1210, 2011.

P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

P. Royston and W. Sauerbrei. Interactions between treatment and continuous covariates: A step toward individualizing therapy. *Journal of Clinical Oncology*, 26(9):1397–99, 2008.

W. Sauerbrei, P. Royston, and K. Zapien. Detecting an interaction between treatment and a continuous covariate: A comparison of two approaches. *Computational Statistics and Data Analysis*, 51(8):4054–63, 2007.

S Solomon, M Rice, K Jablonski, J Jose, M Domanski, M Sabatine, B Gersh, J Rouleau, M Pfeffer, E Braunwald, and (PEACE) Investigators. Renal function and effectiveness of angiotensin-converting enzyme inhibitor therapy in patients with chronic stable coronary disease in the prevention of events with ace inhibition (peace) trial. *Circulation*, 114:26–31, 2006.

X. Su, T. Zhou, X. Yan, F. Fan, and S. Yang. Interaction trees with censored survival data. *The International Journal of Biostatistics*, 4(1):Article 2, 2008.

L. Tian and R. Tibshirani. Adaptive index models for marker-based risk stratification. *Biostatistics*, page 10.1093/biostatistics/kxq047, 2010.

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

R. Tibshirani, T. Hastie, B. Narasimhan, and C. Gilbert. Prediction analysis for microarrays (pam) software. 2003. $/home/tibs/public_html/PAM$.

S. Van De Geer. High-dimensional gneralized linear models and lasso. *Annals of Statistics*, 36:614–645, 2008.

S. Van De Geer and P Buhlmann. On the conditions used to prove oracle results for lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

C. Zhang and J Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

Y. Zhao, D. Zeng, A. Rush, and M. Kosorok. Estimating individualized treatement rules using outcome weighted learning. *Journal of the American Statistical Association*, 107 (499):1106–1118, 2012.

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

H. Zou and T. Hastie. Regularization and variable selection via elastic net. *Journal of Royal Statistical Society. B*, 67:301–320, 2005.

6 Appendix

6.1 Justification of the Objective Function Based on the Working Model

Under the linear working model for continuous responses, we have

$$\mathrm{E}\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=1\} = \frac{1}{2}\left[\mathrm{E}\{(Y^{(1)})^2|\mathbf{Z}\} - 2m_1(\mathbf{Z})f(\mathbf{Z}) + f(\mathbf{Z})^2\right]$$

and

$$\mathrm{E}\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=-1\} = \frac{1}{2}\left[\mathrm{E}\{(Y^{(-1)})^2|\mathbf{Z}\} + 2m_{-1}(\mathbf{Z})f(\mathbf{Z}) + f(\mathbf{Z})^2\right],$$

where $m_t(\mathbf{z}) = \mathrm{E}(Y^{(t)}|\mathbf{Z}=\mathbf{z})$ for $t = 1$ and -1. Therefore

$$\begin{aligned}
\mathcal{L}(f) =& \mathrm{E}\{l(Y, f(\mathbf{Z})T)\} \\
=& \mathrm{E}_{\mathbf{Z}}\left[\frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=1\} + \frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=-1\}\right] \\
=& \mathrm{E}_{\mathbf{Z}}\left(\left[\frac{1}{2}\{m_1(\mathbf{Z}) - m_{-1}(\mathbf{Z})\} - f(\mathbf{Z})\right]^2\right) + \text{constant}.
\end{aligned}$$

The minimizer of this objective function is

$$f^*(\mathbf{z}) = \frac{1}{2}\{m_1(\mathbf{z}) - m_{-1}(\mathbf{z})\} = \frac{1}{2}\Delta(\mathbf{z})$$

for all $\mathbf{z} \in$ Support of $\mathbf{Z}$.

Under the logistic working model for binary responses, we have

$$\mathrm{E}\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T = 1\} = m_1(\mathbf{Z})f(\mathbf{Z}) - \log(1 + e^{f(\mathbf{Z})}),$$

and

$$\mathrm{E}\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T = -1\} = -m_{-1}(\mathbf{Z})f(\mathbf{Z}) - \log(1 + e^{-f(\mathbf{Z})}).$$

Thus

$$\begin{aligned}
\mathcal{L}(f) =& \mathrm{E}\{l(Y, f(\mathbf{Z})T)\} \\
=& \mathrm{E}_{\mathbf{Z}}\left[\frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T = 1\} + \frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T = -1\}\right] \\
=& \frac{1}{2}\mathrm{E}_{\mathbf{Z}}\left[\Delta(\mathbf{Z})f(\mathbf{Z}) - \log(1 + e^{f(\mathbf{Z})}) - \log(1 + e^{-f(\mathbf{Z})})\right].
\end{aligned}$$

Therefore

$$\frac{\partial \mathcal{L}(f)}{\partial f} = \frac{1}{2}\mathrm{E}_{\mathbf{Z}}\left[\Delta(\mathbf{Z}) - \frac{1 - e^{f(\mathbf{Z})}}{1 + e^{f(\mathbf{Z})}}\right],$$

which implies that the minimizer of $\mathcal{L}(f)$ is

$$f^*(\mathbf{z}) = \log\frac{1 - \Delta(\mathbf{z})}{1 + \Delta(\mathbf{z})}$$

for all $\mathbf{z} \in$ Support of $\mathbf{Z}$ or equivalently

$$\Delta(\mathbf{z}) = \frac{1 - e^{f^*(\mathbf{z})}}{1 + e^{f^*(\mathbf{z})}}.$$

Alternatively, under the logistic working model with binary responses, we may focus on the objective function

$$\tilde{l}(Y, f(\mathbf{Z})T) = (1-Y)f(\mathbf{Z})T + Ye^{-f(\mathbf{Z})T}.$$

Therefore

$$\mathrm{E}\{\tilde{l}(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=1\} = \{1-m_1(\mathbf{Z})\}f(\mathbf{Z}) + m_1(\mathbf{Z})e^{-f(\mathbf{Z})},$$

and

$$\mathrm{E}\{\tilde{l}(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=-1\} = -\{1-m_{-1}(\mathbf{Z})\}f(\mathbf{Z}) + m_{-1}(\mathbf{Z})e^{f(\mathbf{Z})}.$$

Thus

$$\begin{aligned}
\mathcal{L}(f) =& \mathrm{E}\{\tilde{l}(Y, f(\mathbf{Z})T)\} \\
=& \mathrm{E}_{\mathbf{Z}}\left[\frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=1\} + \frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T=-1\}\right] \\
=& \mathrm{E}_{\mathbf{Z}}\left[\frac{1}{2}\{m_{-1}(\mathbf{Z}) - m_1(\mathbf{Z})\}f(\mathbf{Z}) + \frac{1}{2}m_1(\mathbf{Z})e^{-f(\mathbf{Z})} + m_{-1}(\mathbf{Z})e^{f(\mathbf{Z})}\right].
\end{aligned}$$

Therefore

$$\frac{\partial\mathcal{L}(f)}{\partial f} = \frac{1}{2}\mathrm{E}_{\mathbf{Z}}\left[\{m_{-1}(\mathbf{Z}) - m_1(\mathbf{Z})\} - m_1(\mathbf{Z})e^{-f(\mathbf{Z})} + m_{-1}(\mathbf{Z})e^{f(\mathbf{Z})}\right]$$

which implies that the minimizer of $\mathcal{L}(f)$ is

$$f^*(\mathbf{z}) = \log\frac{m_1(\mathbf{z})}{m_{-1}(\mathbf{z})}$$

for all $\mathbf{z} \in$ Support of $\mathbf{Z}$.

Under the Cox working model for survival outcomes, we have

$$\begin{aligned}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T\} =& \mathrm{E}_Y\left(\int_0^\tau \left[Tf(\mathbf{Z}) - \log\{\mathrm{E}(e^{Tf(\mathbf{Z})}I(X \geq u))\}\right] dN(u)|\mathbf{Z}, T\right) \\ =& \int_0^\tau \left[Tf(\mathbf{Z}) - \log\{\mathrm{E}(e^{Tf(\mathbf{Z})}I(X \geq u))\}\right] \mathrm{E}\left\{I(X \geq u)|\mathbf{Z}, T\right\} \lambda_T(u; \mathbf{Z})du\end{aligned}$$

where $\lambda_t(u; \mathbf{Z})$ is the hazard function for $\tilde{X}^{(t)}$ given $\mathbf{Z}$ for $t = 1/-1$. Since

$$\mathcal{L}(f) = \mathrm{E}_{\mathbf{Z}}\left[\frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T = 1\} + \frac{1}{2}\mathrm{E}_Y\{l(Y, f(\mathbf{Z})T)|\mathbf{Z}, T = -1\}\right],$$

$$\begin{aligned}\frac{\partial\mathcal{L}(f)}{\partial f} =& \frac{1}{2}\mathrm{E}\int_0^\tau \Big\{ I(X^{(1)} \geq u)\lambda_1(u; \mathbf{Z}) - I(X^{(-1)} \geq u)\lambda_{-1}(u; \mathbf{Z}) \\ & - e^{f(\mathbf{Z})}I(X^{(1)} \geq u)\lambda(u; f) + e^{-f(\mathbf{Z})}I(X^{(-1)} \geq u)\lambda(u; f) \Big\} du,\end{aligned}$$

where $X^{(j)} = \tilde{X}^{(j)} \wedge C^{(j)}$, $C^{(j)}$ is the censoring time if the patient is assigned to the group j,

$$\lambda(t; f) = \frac{\mathrm{E}[I(X \geq t)\{\lambda_T(t; \mathbf{Z})\}]}{\mathrm{E}\{e^{Tf(\mathbf{Z})}I(X \geq t)\}}.$$

Setting the derivative at zero, the minimizer $f^*(\mathbf{z})$ satisfies

$$\begin{aligned}& e^{f^*(\mathbf{z})}\mathrm{E}\{\Lambda^*(\tilde{X}^{(1)})|\mathbf{Z} = \mathbf{z}\} - e^{-f^*(\mathbf{z})}\mathrm{E}\{\Lambda^*(\tilde{X}^{(-1)})|\mathbf{Z} = \mathbf{z}\} \\ =& \mathrm{Prob}(\delta = 1|T = 1, \mathbf{Z} = \mathbf{z}) - \mathrm{Prob}(\delta = 1|T = -1, \mathbf{Z} = \mathbf{z})\end{aligned}$$

for all $\mathbf{z} \in$ Support of $\mathbf{Z}$, where $\Lambda^*(t) = \int_0^\infty \int_0^t \lambda(u \wedge c; f^*)f_C(c)dudc$ is an increasing function of t. Here, we assume that the censoring time and $(\mathbf{Z}, T)$ are independent and $f_C(\cdot)$ is the density function of the censoring time. This assumption is reasonable when the censoring is due to administrative reasons. Furthermore, when censoring rates in two arms are close to

each other for all given $\mathbf{z}$, i.e., $\text{Prob}(\delta = 1|T = 1, \mathbf{Z} = \mathbf{z}) \approx \text{Prob}(\delta = 1|T = -1, \mathbf{Z} = \mathbf{z})$,

$$f^*(z) \approx -\frac{1}{2}\log\left[\frac{E\{\Lambda^*(\tilde{X}^{(1)})|\mathbf{Z} = \mathbf{z}\}}{E\{\Lambda^*(\tilde{X}^{(-1)})|\mathbf{Z} = \mathbf{z}\}}\right].$$

6.2 Justification of the Optimal $a_0(\mathbf{z})$ in the Efficiency Augmentation

Let $S(y, \mathbf{w}^*, \boldsymbol{\gamma})$ be the derivative of the objective function $l(y, \boldsymbol{\gamma}'\mathbf{w}^*)$ with respect to $\boldsymbol{\gamma}$. $\hat{\boldsymbol{\gamma}}$ is the root of an estimating equation

$$Q(\boldsymbol{\gamma}) = N^{-1}\sum_{i=1}^{N} S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}) = 0.$$

Similarly, the augmented estimator $\hat{\boldsymbol{\gamma}}_a$ can be viewed as the root of the estimating equation

$$Q_a(\boldsymbol{\gamma}) = N^{-1}\sum_{i=1}^{N}\{S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}) - T_i \cdot \mathbf{a}(\mathbf{Z}_i)\} = 0.$$

Since $E\{T_i \cdot \mathbf{a}(\mathbf{Z}_i)\} = 0$ due to randomization, the solution of the augmented estimating equation always converges to $\boldsymbol{\gamma}^*$ in probability. It is straightforward to show that

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* = N^{-1}A_0^{-1}\sum_{i=1}^{N} S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}^*) + o_P(N^{-1})$$

and

$$\hat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}^* = N^{-1}A_0^{-1}\sum_{i=1}^{N}\{S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}^*) - T_i\mathbf{a}(\mathbf{Z}_i)\} + o_P(N^{-1}),$$

where A_0 is the derivative of $\text{E}\{S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma})\}$ with respect to $\boldsymbol{\gamma}$ at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$. Selecting the optimal $\mathbf{a}(\mathbf{z})$ is equivalent to minimizing the variance of $\{S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}^*) - T_i\mathbf{a}(\mathbf{Z}_i)\}$. Noting

that

$$\mathrm{E}\left[\{S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}^*) - T_i\mathbf{a}(\mathbf{Z}_i)\}^{\otimes 2}\right] = \mathrm{E}\left[\{S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}^*) - T_i\mathbf{a}_0(\mathbf{Z}_i)\}^{\otimes 2}\right] + \mathrm{E}[\{\mathbf{a}(\mathbf{Z}_i) - \mathbf{a}_0(\mathbf{Z}_i)\}^{\otimes 2}],$$

where $\mathbf{a}_0(\mathbf{z})$ satisfies the equation

$$E\left[\{S(Y, \mathbf{W}^*, \boldsymbol{\gamma}^*) - T\mathbf{a}_0(\mathbf{Z})\}T\eta(\mathbf{Z})\right] = 0$$

for any function $\eta(\cdot)$, $\mathbf{a}_0(\cdot)$ is the optimal augmentation term minimizing the variance of $\hat{\boldsymbol{\gamma}}_a$. Since $\mathbf{a}_0(\cdot)$ is the root of the equation

$$\mathrm{E}\left[\{S(Y, \mathbf{W}^*, \boldsymbol{\gamma}^*) - T\mathbf{a}_0(\mathbf{Z})\}'T \;\middle|\; \mathbf{Z} = \mathbf{z}\right] = 0,$$

$$\mathbf{a}_0(\mathbf{z}) = \frac{1}{2}\left[\mathrm{E}\{S(Y, \mathbf{W}(\mathbf{z})/2, \boldsymbol{\gamma}^*)|\mathbf{Z} = \mathbf{z}, T = 1\} - \mathrm{E}\{S(Y, -\mathbf{W}(\mathbf{z})/2, \boldsymbol{\gamma}^*)|\mathbf{Z} = \mathbf{z}, T = -1\}\right].$$

For continuous responses,

$$S(Y, \mathbf{W}^*, \boldsymbol{\gamma}) = -\frac{1}{2}T\mathbf{W}(\mathbf{Z})\left\{Y - \frac{1}{2}T\mathbf{W}(\mathbf{Z})'\boldsymbol{\gamma}\right\}$$

and

$$\begin{aligned} a_0(\mathbf{z}) =& \frac{1}{2}\left(\mathrm{E}[-\mathbf{W}(\mathbf{z})\{Y - \mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2\}/2|T = 1, \mathbf{Z} = \mathbf{z}] - \mathrm{E}[\mathbf{W}(\mathbf{z})\{Y + \mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2\}/2|T = -1, \mathbf{Z} = \mathbf{z}]\right) \\ =& -\mathbf{W}(\mathbf{z})\left\{\frac{1}{4}\mathrm{E}(Y|T = 1, \mathbf{Z} = \mathbf{z}) + \frac{1}{4}\mathrm{E}(Y|T = -1, \mathbf{Z} = \mathbf{z})\right\} \\ =& -\frac{1}{2}\mathbf{W}(\mathbf{z})\mathrm{E}(Y|\mathbf{Z} = \mathbf{z}). \end{aligned}$$

For binary responses,

$$S(Y, \mathbf{W}^*, \boldsymbol{\gamma}) = -\frac{1}{2}\mathbf{W}(\mathbf{Z})T\left\{Y - \frac{e^{T\mathbf{W}(\mathbf{Z})'\boldsymbol{\gamma}/2}}{1+e^{T\mathbf{W}(\mathbf{Z})'\boldsymbol{\gamma}/2}}\right\}$$

and

$$\begin{aligned}
a_0(\mathbf{z}) =& -\frac{1}{4}\mathbf{W}(\mathbf{z})\left[\mathrm{E}\left\{Y - \frac{e^{\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}}{1+e^{\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}} \;\middle|\; T=1, \mathbf{Z}=\mathbf{z}\right\} + \mathrm{E}\left\{Y - \frac{e^{-\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}}{1+e^{-\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}} \;\middle|\; T=-1, \mathbf{Z}=\mathbf{z}\right\}\right] \\
=& -\frac{1}{4}\mathbf{W}(\mathbf{z})\left\{\mathrm{E}(Y|T=1, \mathbf{Z}=\mathbf{z}) + \mathrm{E}(Y|T=-1, \mathbf{Z}=\mathbf{z}) - \left(\frac{e^{\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}}{1+e^{\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}} + \frac{e^{-\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}}{1+e^{-\mathbf{W}(\mathbf{z})'\boldsymbol{\gamma}^*/2}}\right)\right\} \\
=& -\frac{1}{2}\mathbf{W}(\mathbf{z})\left\{\mathrm{E}(Y|\mathbf{Z}=\mathbf{z}) - \frac{1}{2}\right\}.
\end{aligned}$$

For survival responses, the estimating equation based on the partial likelihood function is asymptotically equivalent to the estimating equation $N^{-1}\sum_{i=1}^{N} S(Y_i, \mathbf{W}_i^*, \boldsymbol{\gamma}) = 0$, where

$$S(Y, \mathbf{W}^*, \boldsymbol{\gamma}) = -\int_0^\tau \left[\mathbf{W}^* - \mathbf{R}(u;\boldsymbol{\gamma})\right] M(du, \mathbf{W}^*, \boldsymbol{\gamma}).$$

Thus

$$\mathbf{a}_0(\mathbf{z}) = -\frac{1}{2}\left[\frac{1}{2}\mathbf{W}(\mathbf{z})\left\{G_1(\tau;\mathbf{z}) + G_2(\tau;\mathbf{z})\right\} - \int_0^\tau \mathbf{R}(u)\{G_1(du;\mathbf{z}) - G_2(du;\mathbf{z})\}\right].$$

6.3 The Lasso Algorithm in the Efficiency Augmentation

In general, the augmentation term is in the form of $\mathbf{a}_0(\mathbf{Z}_i) = \mathbf{W}(\mathbf{Z}_i)'\hat{r}(\mathbf{Z}_i)$, where $\hat{r}(\mathbf{Z}_i)$ is a simple scalar. The lasso regularized objective function can be written as

$$\frac{1}{N}\sum_{i=1}^{N}\{l(Y_i, \boldsymbol{\gamma}'\mathbf{W}_i^*) - \boldsymbol{\gamma}'\mathbf{W}_i^*\hat{r}(\mathbf{Z}_i)\} + \lambda|\boldsymbol{\gamma}|.$$

In general, this lasso problem can be solved iteratively. For example, when $l(\cdot)$ is the log-likelihood function of the logistic regression model, we may update $\hat{\boldsymbol{\gamma}}$ iteratively by solving the standard OLS-lasso problem

$$\frac{1}{N}\sum_{i=1}^{N}\hat{w}_i(\hat{z}_i - \boldsymbol{\gamma}'\mathbf{W}_i^*)^2 + \lambda\|\boldsymbol{\gamma}\|_1,$$

where

$$\hat{z}_i = \hat{\boldsymbol{\gamma}}'\mathbf{W}_i^* + \hat{w}_i^{-1}\{Y_i - \hat{p}_i - \hat{r}(\mathbf{Z}_i)\}, \qquad \hat{w}_i = \hat{p}_i(1-\hat{p}_i),$$

$\hat{\boldsymbol{\gamma}}$ is the current estimator for $\boldsymbol{\gamma}$ and

$$\hat{p}_i = \frac{\exp\{\hat{\boldsymbol{\gamma}}'\mathbf{W}_i^*\}}{1+\exp\{\hat{\boldsymbol{\gamma}}'\mathbf{W}_i^*\}}.$$

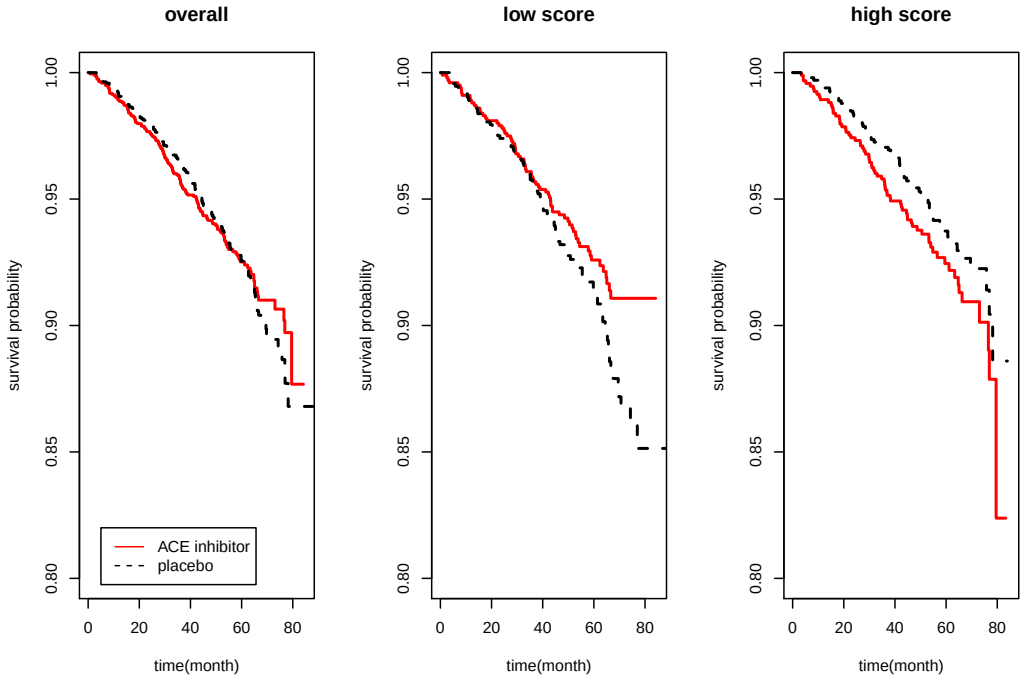

Figure 1: Example of the modified covariate approach, applied to patients with stable coronary artery disease and normal or slightly reduced left ventricular function, who were randomly given ACE inhibitor or placebo in a randomized trial. Our procedure constructed a score based on baseline covariates to detect covariates-treatment interactions. The numerical score was constructed on a training set, and then categorized into low and high. The panels show the survival curves for a separate validation set, overall and stratified by the score.

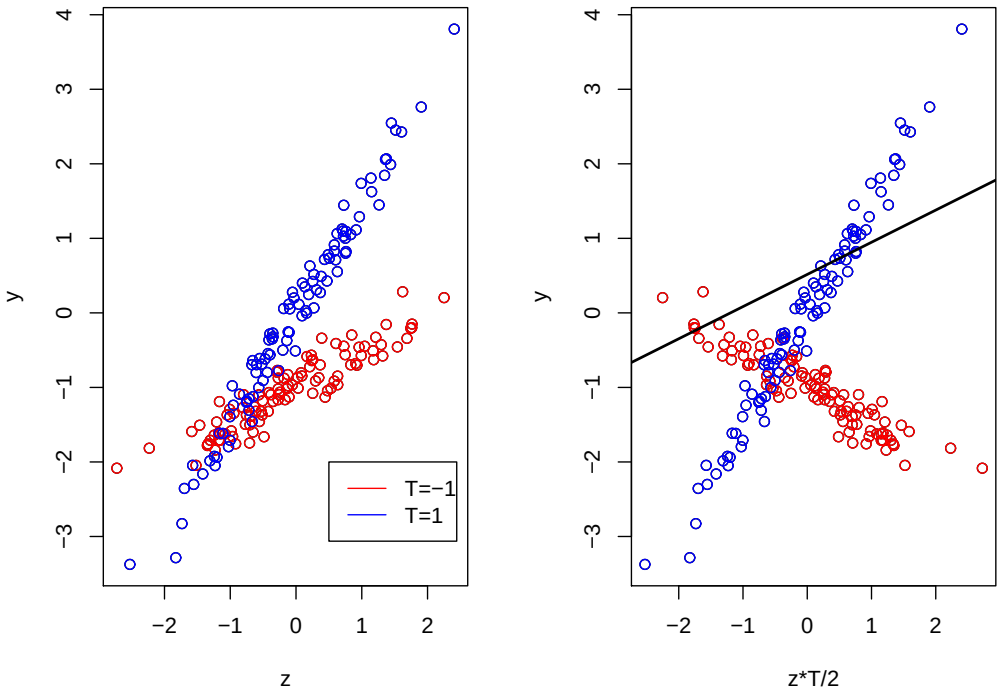

Figure 2: Example of the modified covariate approach. The raw data is shown on the left, consisting of a single covariate Z and a treatment $T = -1$ or 1. The treatment-covariate interaction has a slope γ approximately being 1. On the right panel, we have plotted the response against $Z \cdot T/2$. The regression line computed in the right panel estimates the treatment effect for each given value of covariate Z.

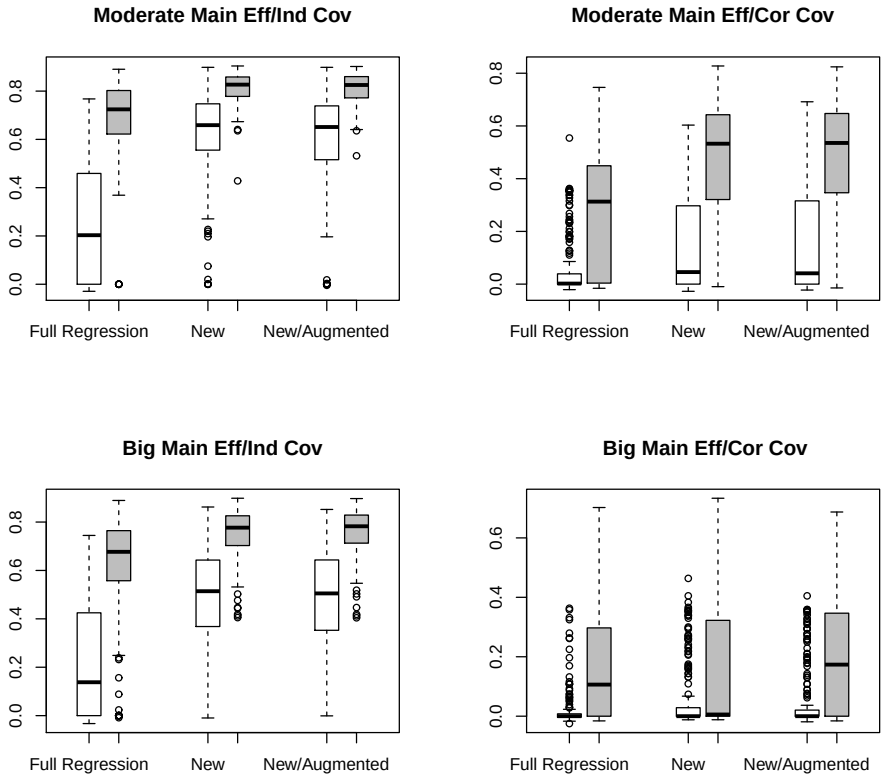

Figure 3: Boxplots for the correlation coefficients between the estimated score and true treatment effect with three different methods applied to continuous outcomes. The empty and filled boxes represent high and low dimensional ($p = 1000$ and $p = 50$) cases, respectively. Left upper panel: moderate main effect and independent covariates; right upper panel: moderate main effect and correlated covariates; left lower panel: big main effect and independent covariates; right lower panel: big main effect and correlated covariates.

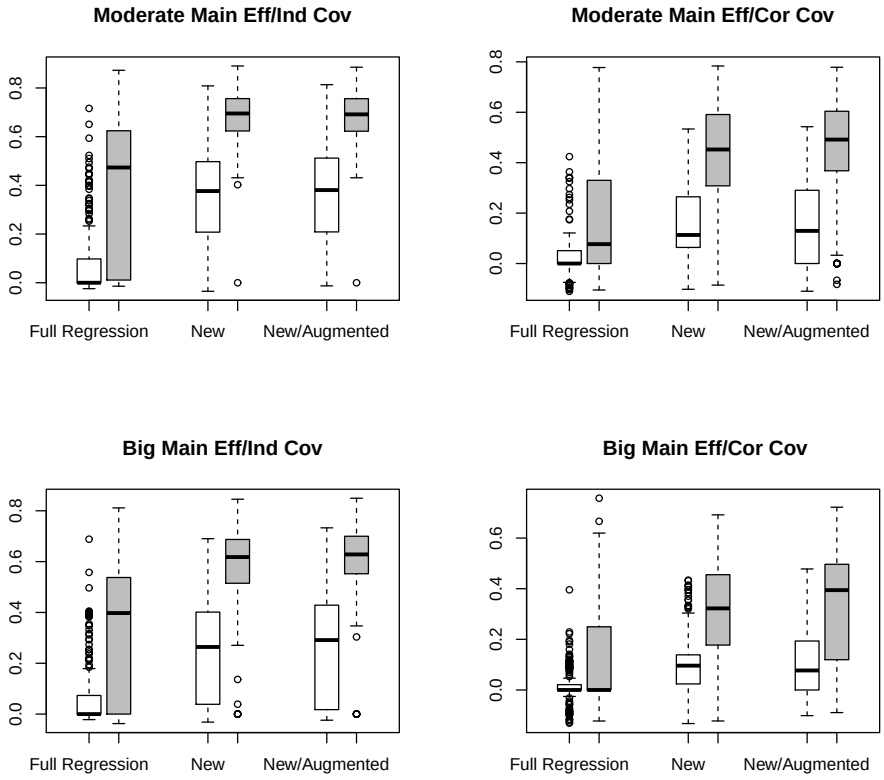

Figure 4: Boxplots for the correlation coefficients between the estimated score and true treatment effect with three different methods applied to binary outcomes. The empty and filled boxes represent high and low dimensional ($p = 1000$ and $p = 50$) cases, respectively. Left upper panel: moderate main effect and independent covariates; right upper panel: moderate main effect and correlated covariates; left lower panel: big main effect and independent covariates; right lower panel: big main effect and correlated covariates.

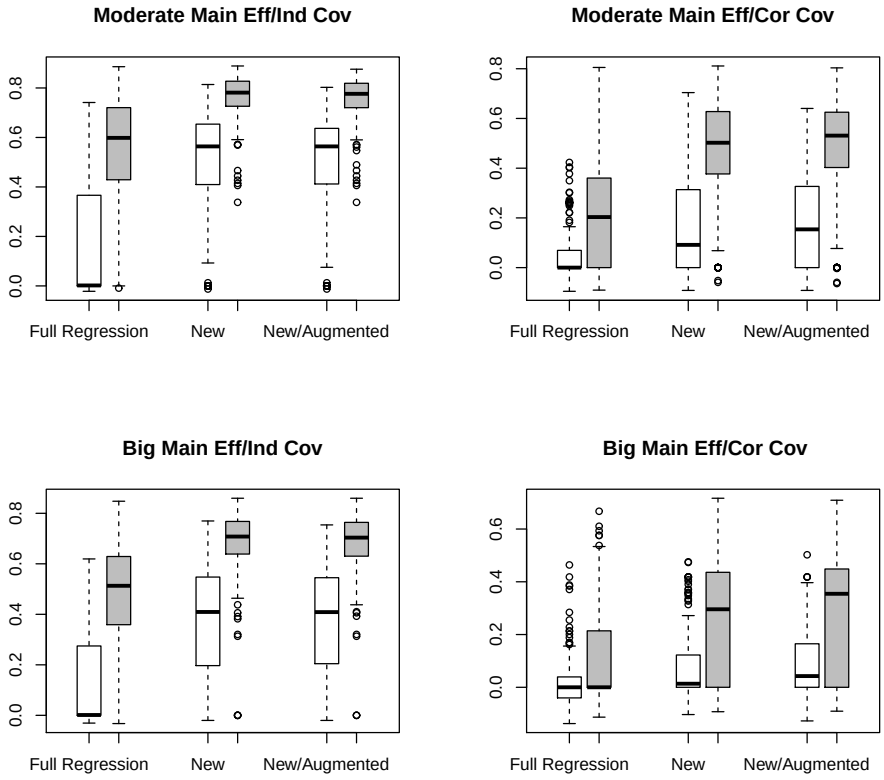

Figure 5: Boxplots for the correlation coefficients between the estimated score and true treatment effect with three different methods applied to survival outcomes. The empty and filled boxes represent high and low dimensional ($p = 1000$ and $p = 50$) cases, respectively. Left upper panel: moderate main effect and independent covariates; right upper panel: moderate main effect and correlated covariates; left lower panel: big main effect and independent covariates; right lower panel: big main effect and correlated covariates.

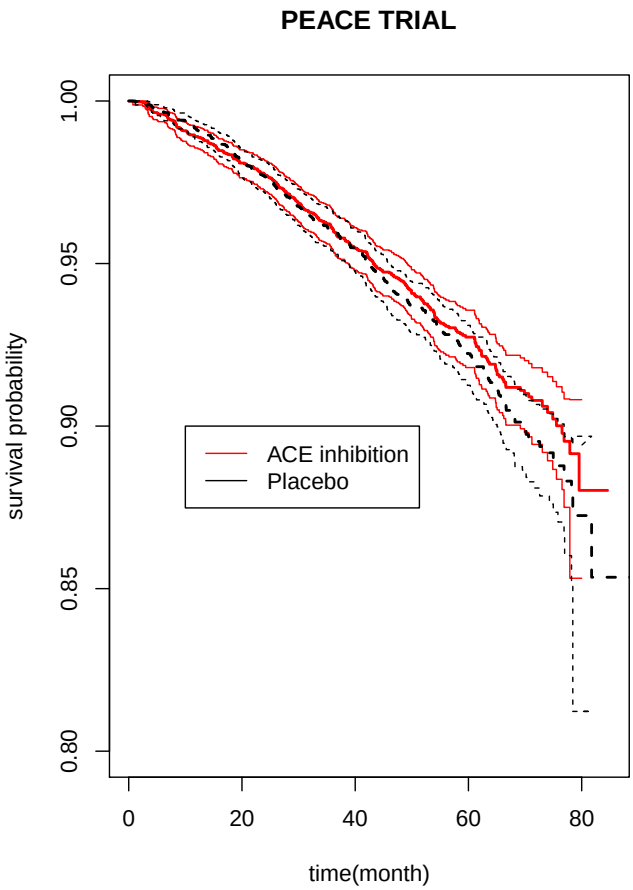

Figure 6: Survival functions of the ACE inhibitor and placebo arms (7865 patients): red line, the ACE inhibitor arm; black line, the placebo arm; thin line, the point-wise 95% confidence limits.

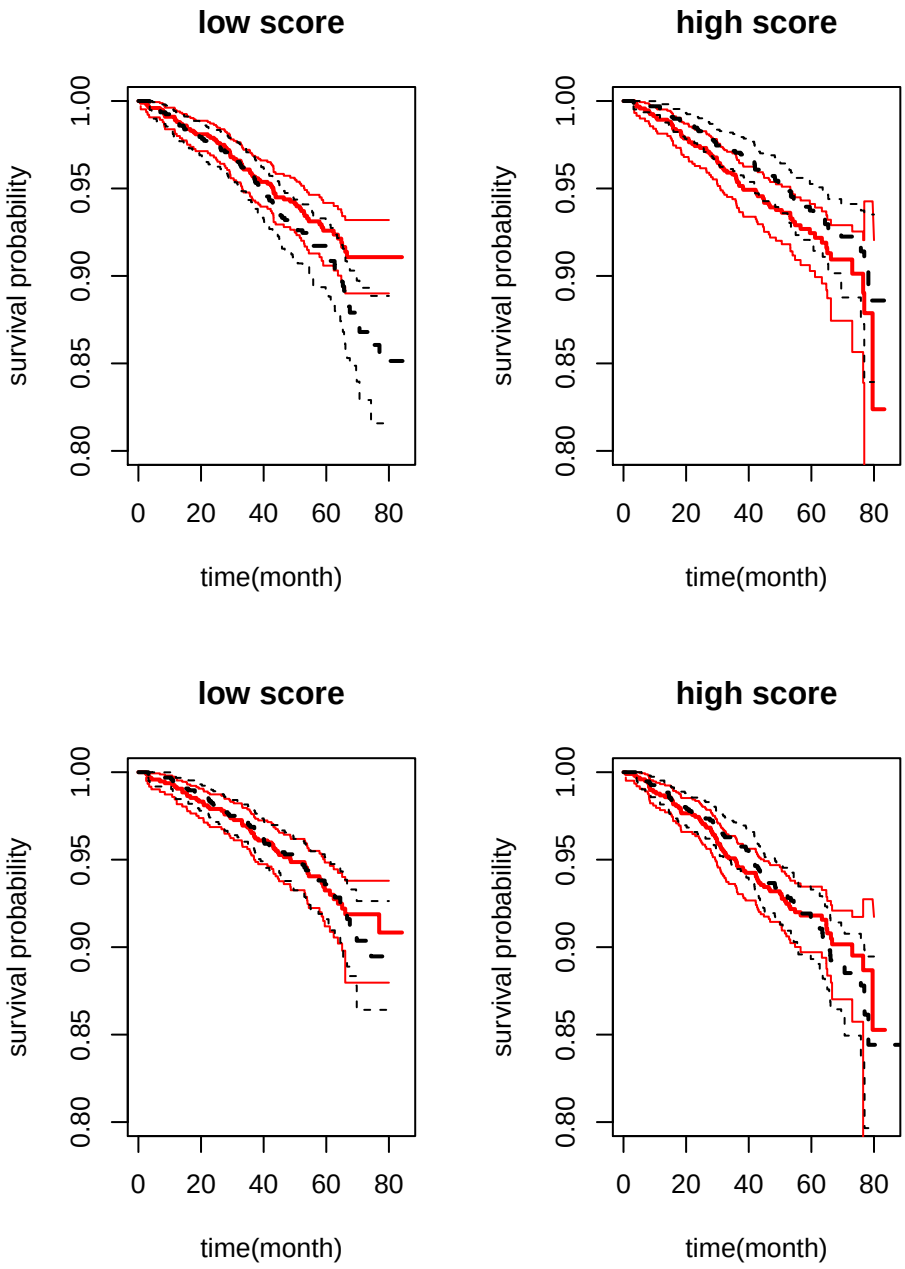

Figure 7: Survival functions of the ACE inhibitor and placebo arms stratified by the estimated score in the validation set: red line, the ACE inhibitor arm; black line, the placebo arm; thin line, the point-wise 95% confidence limits. Upper panels: the score is based on the "new" method; lower panel: the score is based on the "full regression" method.

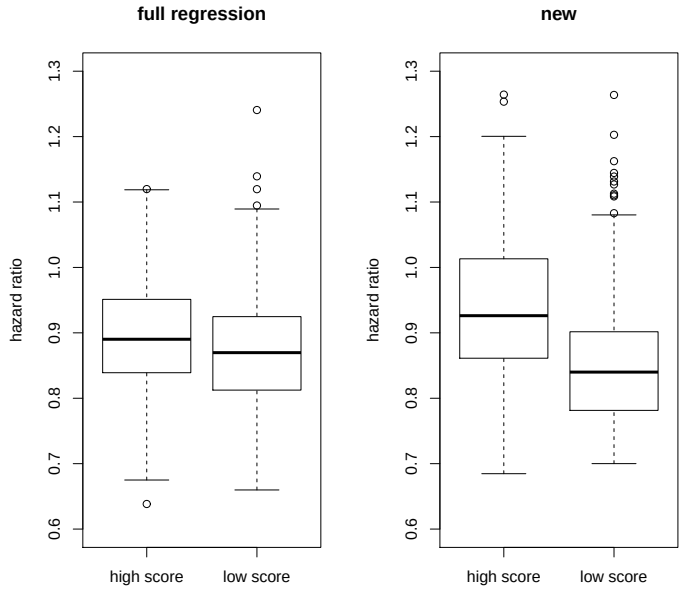

Figure 8: Boxplots for hazard ratios in high and low risk groups based on 500 random splits of the PEACE data. A big difference between the two groups represents high quality of the constructed score in stratifying patients according to the individualized treatment effect.

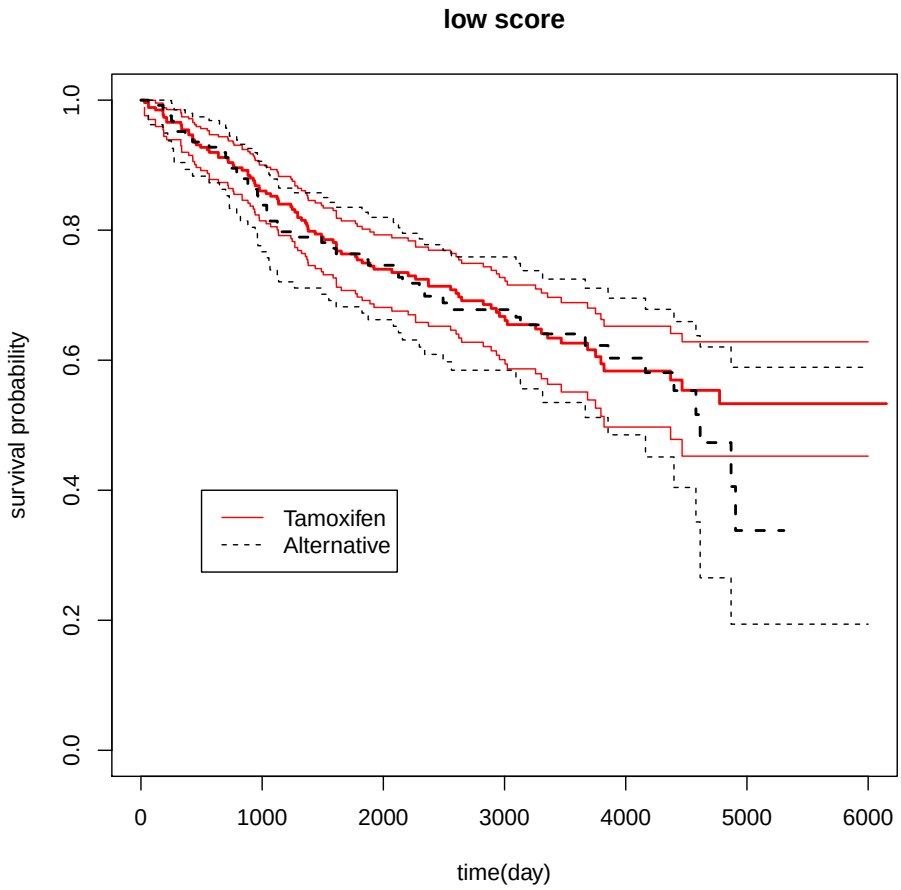

Figure 9: Survival functions of the Tamoxifen and alternative treatment groups: red line, the Tamoxifen group; black line, the alternative treatment arm; thin line, the point-wise 95% confidence limits.

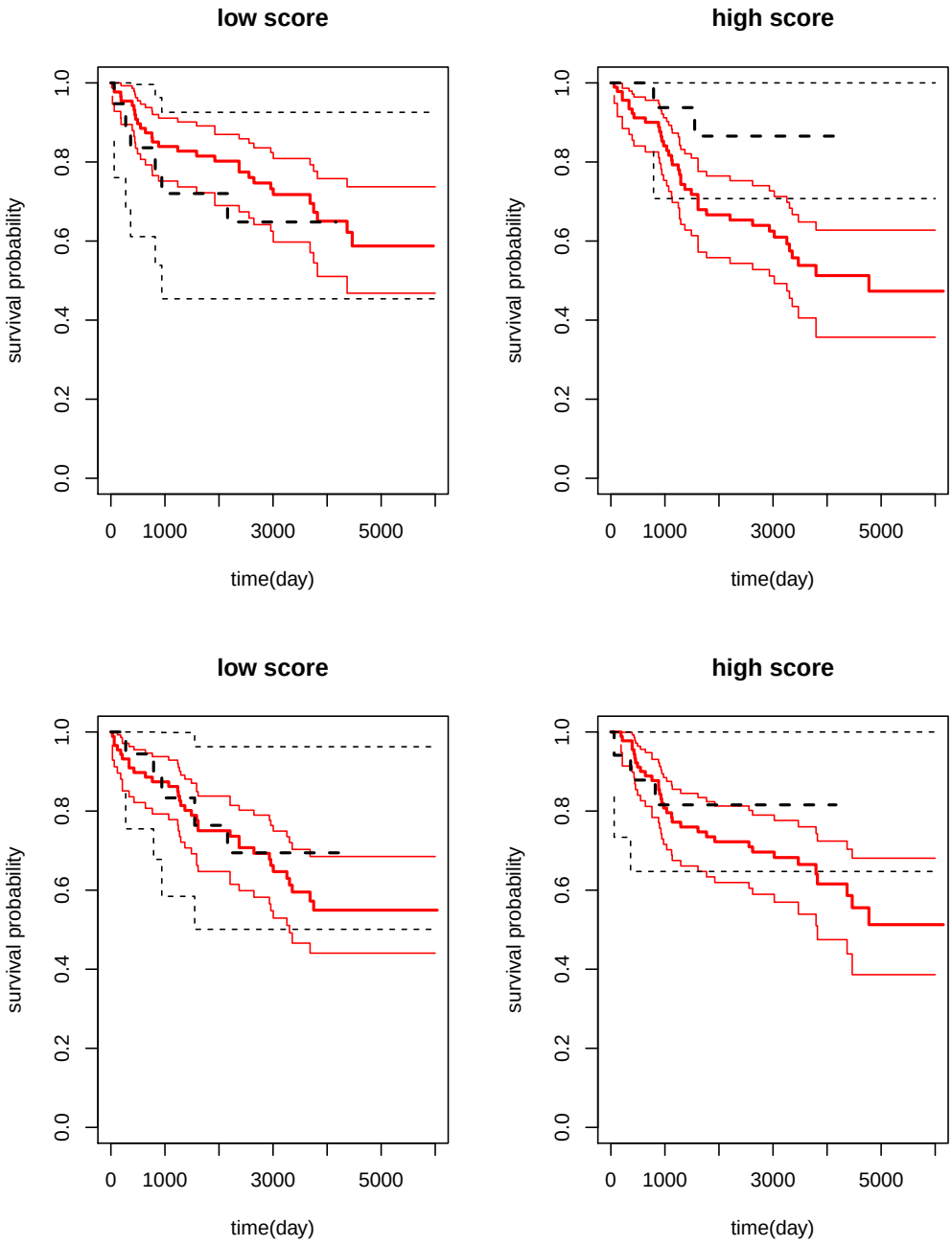

Figure 10: Survival functions of the Tamoxifen and alternative treatment groups stratified by the estimated score in the validation set: red line, the Tamoxifen treatment group; black line, the alternative treatment group; thin line, the point-wise 95% confidence limits. Upper panels: the score is based on the "new" method; lower panel: the score is based on "full regression" method.