HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2016 May 10.

Published in final edited form as:

Stat Med. 2015 May 10; 34(10): 1645–1658. doi:10.1002/sim.6433.

Penalized regression procedures for variable selection in the potential outcomes framework

Debashis Ghosh[1], **Yeying Zhu**[2], and **Donna L. Coffman**[3]

[1]Department of Biostatistics and Informatics, Colorado School of Public Health Aurora, CO, 80045, USA

[2]Department of Statistics and Actuarial Science, University of Waterloo Waterloo, ON, N2L 3G1, Canada

[3]Methodology Center, Penn State University, University Park, PA, 16802, USA

Abstract

A recent topic of much interest in causal inference is model selection. In this article, we describe a framework in which to consider penalized regression approaches to variable selection for causal effects. The framework leads to a simple 'impute, then select' class of procedures that is agnostic to the type of imputation algorithm as well as penalized regression used. It also clarifies how model selection involves a multivariate regression model for causal inference problems, and that these methods can be applied for identifying subgroups in which treatment effects are homogeneous. Analogies and links with the literature on machine learning methods, missing data and imputation are drawn. A difference LASSO algorithm is defined, along with its multiple imputation analogues. The procedures are illustrated using a well-known right heart catheterization dataset.

Keywords

Average causal effect; Counterfactual; Imputed data; L_1 penalty; Treatment heterogeneity

1. Introduction

In many medical and scientific studies, investigators are interested in making causal statements about the effect of a treatment on outcomes. For a well-designed randomized study, we assume that any covariates that may influence the outcome have the same distribution among different treatment groups. Consequently, the treatment is the only factor that may cause differences in the outcome. However, in an observational study, if the treatment assignment is not randomized, there usually exists a set of confounders that may influence both the outcome and the treatment assignment. In this case, any causal inference failing to account for the confounders will lead to biased estimates of the treatment effect.

A very popular framework for causal effects is the potential outcomes model [1, 2]. This framework formulates counterfactual random variables that represent the outcome variable under the hypothetical treatments of interest for each individual. Causal estimands are then defined based on the contrasts between the within-individual counterfactuals. Because the

intervention of interest is typically not randomized in observational studies, further modelling is required for proper estimation of causal effects. Many approaches exist in the literature, including g-estimation [3], matching [4], double-robust estimation [5] and inverse probability weighted estimation [6]. One key quantity in causal modelling is that of propensity scores [7]. This quantity is defined as the probability of the treatment given covariates. Based on the estimated propensity scores, a variety of modelling strategies can be used to estimate causal effects; a good summary of them can be found in [8].

A question of much recent interest is that of how to select variables to use for the estimation of causal effects. This is keeping in line with the interest in penalized regression approaches in the statistical literature (e.g., [9, 10]). Many approaches for variable selection in causal inference have been described recently. A proposal from Hirano and Imbens [11] is to consider predictors based on univariate tests in both the propensity score model as well as the mean outcome models. Simulation evidence presenting the importance of variable selection was provided in [12]. Model averaging approaches for causal effects have been advocated by several authors [13, 14, 15]. An algorithmic approach in which cross-validation is used to select the optimal model for causal inference has been developed by Brookhart and van der Laan [16]; this is also related to the general targeted learning framework that has been summarized in the recent monograph by van der Laan and Rose [17]. For the causal graphical modelling framework of Pearl [18], which is based on directed acyclic graphs, Bühlmann et al. [19] proposed the use of a high-dimensional screening technique for variable selection.

The approach taken in this paper is based on the original LASSO proposal from [9] and explicitly makes use of the predictive nature of the causal inference problem. By prediction, we mean that we treat the problem of estimating causal effects as a missing data problem and use the predictive distribution of the missing data given the observed data to ”fill in” the missing potential outcomes. This naturally lends itself to the use of techniques from the multiple imputation literature [20]. The prediction point of view is discussed in §2.3. This will lead us to an application of what has been termed the predictive LASSO [21] for performing variable selection in the potential outcomes model. The structure of this paper is as follows. In Section 2, we describe the observed data and review the potential outcomes framework. We also describe the conceptual flaws in applying off-the-shelf variable selection procedures to attempt to perform proper causal inference. The concept of predictive LASSO, as discussed in [21], is described. Section 3 features adaptation of this idea to the causal inference problem. It is seen there that effectively, the variable selection problem is for a multivariate response variable. This leads to a novel LASSO innovation, the difference LASSO. The difference LASSO and its extensions can be used to determine treatment effect heterogeneity in a given dataset. Extensions to this procedure, inspired by multiple imputation, are described in §3.2. Some numerical examples to illustrate the methodology are given in Section 4. Some discussion concludes Section 5.

2. Preliminaries

2.1 Review of potential outcomes framework

We define T to be the treatment indicator that takes values zero and one. The random variables $\{Y(0), Y(1)\}$ are the potential outcomes for the subject under $T = 0$ and $T = 1$, respectively. What we observe is $Y_i \equiv Y_i(T_i)$ ($i = 1, \ldots, n$), which implies that $Y(0)$ and $Y(1)$ can not be observed simultaneously, i.e. one of them is missing. Two parameters of interest are the average causal effect:

$$ACE=E[Y(1)-Y(0)], \quad (1)$$

and the average causal effect among the treated:

$$ACET=E[Y(1)-Y(0)|T=1]. \quad (2)$$

In an observational study, the vector of covariates $\mathbf{X}$ could be related to both the outcome and the treatment assignment. Since both T and the potential outcomes $\{Y(0),Y(1)\}$ are affected by $\mathbf{X}$, the independence of treatment and the potential outcomes will not hold. This is the situation of confounding and is quite common in epidemiological studies. An important assumption made by Rosenbaum and Rubin [7] that allows for proper causal inference in non-randomized observational studies is called strongly ignorable treatment assignment:

$$T \perp \{Y(0), Y(1)\}|\mathbf{X}. \quad (3)$$

This assumption says that treatment assignment is conditionally independent of the set of potential outcomes given the covariates. In other words, conditioning on the same value of $\mathbf{X}$, we can pretend that the observed outcomes are from a randomized trial. However, conditioning on a p-dimensional vector suffers from the "curse of dimensionality", especially when the dimension is high. Rosenbaum and Rubin [7] further proposed the concept of propensity scores, which is defined as the probability of receiving the treatment given the covariates: $e(\mathbf{X}) \equiv P(T = 1|\mathbf{X})$. Rosenbaum and Rubin [7] show that if (3) holds, the following property is also true:

$$T \perp \{Y(0), Y(1)\}|e(\mathbf{X}). \quad (4)$$

Since $e(\mathbf{X})$ is a scalar quantity, Rosenbaum and Rubin [7] argue that this greatly facilitates the causal inference problem due to a reduction in dimension.

2.2 Variable selection in models: some intuition

In this section, we discuss from an intuitive viewpoint why the problem being addressed cannot be easily handled using existing regression-based variable selection methods. As has been alluded to earlier, there are typically two models being fit: one for the propensity score, and one for the mean outcome. In practice, regression models are fit for propensity score estimation, as well as to the mean outcome model in which the propensity score is accommodated using one of the approaches described in [8].

Figure 1 shows the case of two populations arising from a mixture of normal distributions. Here, there is one confounder. It is distributed as $N(0, 1)$ in the $T = 0$ group and $N(2, 1)$ in the $T = 1$ group. If we were to find the classifier based on X that separates the $T = 1$ and $T = 0$ group, it seen from Figure 1 that there is relatively limited covariate overlap between the two treatment groups, which is a violation of the common support condition [7]. For example, when $X < -2$ or $X > 4$, the estimation of $Y(1) - Y(0)$ will be completely based on extrapolation. Intuitively, the criterion for optimization in the propensity score model does not match up to the ultimate scientific goal, which is "good" estimation of causal effects. This suggests that variable selection for the propensity score model is not sufficient for good causal effect estimation.

Similarly, if we were to perform variable selection of the mean outcome model, then this has problems as well. It represents the scientific model of interest and is intended to identify the causal estimands of interest. Performing variable selection on this model has problems in that different combinations of variables will correspond to different mean outcome models, which naturally will change the scientific question of interest. This discussion is intended to explain why model/variable selection will not be straightforward in the potential outcomes framework.

2.3 Potential outcomes and prediction

With the assumptions described in §2.1, we can characterize the joint distribution of the counterfactuals. An alternative approach would be to begin with regression models for the potential outcomes, such as the structural nested mean models (SNMs, [22]) or marginal structural models (MSMs, [23]). Note that SNMs and MSMs refer to model specification for the potential outcomes and define different target estimands of interest. This is a separate issue from the goal of *estimation*.

Fundamentally, the other issue to realize is that implicitly or explicitly, *predictions* are being made in the modelling of potential outcomes. Coming back to the original potential outcomes framework outlined in §2.1., the complete data consists of $\{Y_i(0), Y_i(1)\}$ for $i = 1, \ldots, n$. This represents the ideal case and would lead to simple calculations for ACE and ACET. However, the observed data have missing values relative to the complete data, and the missingness mechanism is in principle nonignorable, using the terminology of Little and Rubin [24]. The causal assumptions described in §2.1. and in particular, the assumption (3) correspond to missingness at random [24] where the missingness mechanism depends on $\mathbf{X}$. This leads to the following conceptual strategy: using $\mathbf{X}$, impute the missing response variable. Based on the combination of observed and imputed data, one can then compute average causal effects. Suppose we define $R_i = I\{Y_i(0) \text{ missing}\}$, $i = 1, \ldots, n$. Then one can estimate ACE as

$$\hat{ACE} = n^{-1}\left[\sum_{i:R_i=1}(Y_i - Y(0)_i^*) + \sum_{i:R_i=0}(Y(1)_i^* - Y_i)\right], \quad (5)$$

where the asterisks in (5) indicate that the particular potential outcome has been imputed. Another popular methodology is inverse weighted estimation procedures; there, the weights

provide predictions in terms of reweighting the population to one in which the causal effect can be estimated in an unbiased way. Connections between this idea and marginal structural models have been treated in a very simple case in [25]. The reliance on imputation for construction of counterfactuals is intimately tied to the use of the predictive distribution [26] which will inform our modelling strategy in the next section.

2.4. Application of penalized regression methodology to observed outcomes: concepts

As discussed in [8], regression adjustment represents one approach to estimation of causal effects. Suppose we were fitting a regression model for Y on T and $\mathbf{X}$, and suppose that the response variable is continuous. A standard penalized regression to fit would be the LASSO [9], which minimizes the residual sum of squares

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 T_i - \gamma' \mathbf{X}_i)^2 \text{s.t.} |\beta_0| + |\beta_1| + \sum_{j=1}^{p} |\gamma_j| \leq t, \quad (6)$$

where $(\beta_0, \beta_1, \gamma)$ are regression coefficients and $t \geq 0$. Note that the constraint in (6) is on the L_1 norm of the coefficients and leads to situations in which the regression coefficients are estimated to be exactly zero. Variables with zero regression coefficients are deemed to be 'unimportant' predictors. Regression problems with penalties of this form have been the focus of substantial interest in the statistical literature, one reason being that the L_1 penalty induces sparsity in the regression models.

The estimation procedure described so far is simply being applied to the observed data. As alluded to in the previous sections, a problem with the use of (6) for estimation of the causal effects is that it does not take into account the notion that causal effects rely on variables that involve predictions. Given this observation, if we wish to impose an LASSO-style constraint for variable selection, we argue that one should apply it to predictive criterion functions rather than goodness of fit measures from standard regression models such as the sum of squares term in (6). By predictive criterion functions, we are referring to functions of predicted/imputed values of the response. Our general approach is to solve the following optimization problem [21]:

$$\min_{\alpha} \sum_{i=1}^{n} D_i(M_{full}, M_\alpha) + \lambda \sum_{j=1}^{p} |\alpha_j|, \quad (7)$$

where $\lambda > 0$ is a tuning parameter and $D_i(M_{full}, M_\alpha)$ refers to the Kullback-Leibler distance between the predictive distribution of a "full" model relative to those of a model parametrized by the vector α for subject i. When we talk about M here, we are actually alluding to models for the joint distribution of the potential outcomes. The Kullback-Leibler distance is generically defined as

$$D(f, g) = \int \log \left[\frac{g(x)}{f(x)} \right] f(x) dx,$$

where *f* and *g* represent densities of the data. While the optimization problem in (7) is written down in a general form, what Tran et al. [21] show is that in a linear model case, it corresponds to solving a weighted LASSO problem of the form minimizing

$$\sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1 T_i+\hat{\gamma}'\mathbf{X}_i-\beta_0-\beta_1 T_i-\gamma'\mathbf{X}_i)^2 \text{s.t.}|\beta_0|+|\beta_1|+\sum_{j=1}^{p}|\gamma_j|\leq t. \quad (8)$$

where $(\hat{\beta_0}, \hat{\beta_1}, \hat{\gamma})$ are the least squares estimators for the regression coefficients in the usual linear model for the observed data. Implicitly, we are assuming in (8) that the objective function is being evaluated at future design points which are identical to the observed data points. The future design points are also referred to a test set in statistical parlance, which is different from the training data on which the model is built. Tran et al. [21] prove their result using a canonical hierarchical normal model with conjugate priors, and in fact the objective function in (8) arises from the posterior predictive distribution in the model. Further details on the model used in [21] are given in Appendix A.1. Another way to reinterpret their algorithm is as follows:

1. Using a training dataset, fit a linear regression of *Y* on *T* and **X**.
2. Based on the model fitted in step 1., compute predicted/fitted values of *Y* using the test dataset.
3. Solve equation (8) using the test data.

Again, there is an implicit assumption that the training and test datsets will come from the same distribution. In the second step of the algorithm, what is being computed are empirical estimates of the mean of the posterior predictive distribution of *Y* given the observed covariate values. The predictive LASSO occurs in the third step of the algorithm. This reinterpretation also highlights the prediction done in the first two steps and the penalized regression in the third. It also immediately suggests extensions in which the first two steps are substituted with any arbitrary imputation algorithm. Examples of imputation methods include multiply imputed chained equations [27] and IVEWARE [28].

3. Proposed methods

3.1. Adaptation to causal inference

The application of the LASSO method for estimating causal effects requires some care. This is because we fundamentally have a multivariate response for each individual. In particular, we have either ($Y(0)$, $\hat{Y}(1)$) or ($\hat{Y}(0)$, $Y(1)$) depending on which potential outcome is observed. Thus, we wish to perform a LASSO of the multivariate outcome on covariates. We go back to the definition of average causal effect given in (1). This estimand corresponds to the following model:

$$Y_i(1)-Y_i(0)=\tau+\varepsilon_i \quad (9)$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is a random sample from a distribution with mean zero and variance σ^2. The model (9) is a homogeneous treatment effect model for the difference in potential outcomes. However, it might be the case that the causal effect of treatment is in fact heterogeneous so

that (9) is invalid. One can then use the LASSO to identify candidate variables that define subgroups in which the causal effect is homogeneous within the group and heterogeneous among different groups. The selected variables represent the covariates that have what is known as qualitative interaction [32] with treatment. These are interactions in which the treatment effects are heterogeneous among different subgroups defined by the predictor. Following the approach in [21], we now seek to apply the predictive LASSO for the estimation of causal effects.

1. Fit a regression model for Y on T and $\mathbf{X}$ using the training dataset. We will denote the model as M_0.
2. Based on the models fitted in step 1., compute predicted/fitted values of Y using the test dataset to impute the counterfactual or potential outcome described in Section 2. In particular, we will use M_0 to predict $Y(0)$ for subjects with $T = 1$ and predict $Y(1)$ for subjects with $T = 0$.
3. Then, create a univariate response variable for each individual, $\tilde{Y}$, defined to be $Y(1) - \hat{Y}(0)$ if $T = 1$ and $\hat{Y}(1) - Y(0)$ if $T = 0$. A LASSO of $\tilde{Y}$ on the covariates is then performed. Formally, we are solving the following optimization problem:

$$\sum_{i=1}^{n}(\tilde{Y}_i - \beta_0 - \gamma'\mathbf{X}_i)^2 \text{ s.t. } |\beta_0| + \sum_{j=1}^{p}|\gamma_j| \leq t, \quad (10)$$

where (β_0, γ) are unknown regression coefficients, and $t \geq 0$ is a smoothing parameter.

This is equivalent to minimizing:

$$\sum_{i=1}^{n}(\tilde{Y}_i - \beta_0 - \gamma'\mathbf{X}_i)^2 + \lambda(|\beta_0| + \sum_{j=1}^{p}|\gamma_j|), \quad (11)$$

where λ is the smoothing parameter. The procedure we propose here is quite simple in nature and is termed the difference LASSO. There exists attendant software in R, `glmnet` [31], that finds the entire regularization path for the LASSO solution in terms of t, to an optimization problem (10), and we use that here.

In the first two steps, we will use random forests [30] as our prediction algorithm. It belongs to the category of so-called ensemble methods: instead of generating one classification/regression tree, it generates many trees. At each node of a tree, a random subset of the covariates are selected and the node is split based on the best split among the selected covariates. For a testing data point with a covariate vector $\mathbf{X}$, each tree yields an estimated outcome and the final estimate can be calculated by the average value among the trees for continuous outcomes or the majority vote for discrete outcomes. Biau et al. [29] proved the consistency of random forests estimator. In that paper, they also commented that random forests are among the most accurate general-purpose "off-the-shelf" classifiers available.

3.2. Imputation and variable selection

The algorithm described in the previous section can be described as a combination of imputation of the potential outcomes and variable selection. This problem has been addressed in the missing data literature [35, 33, 36]. A comprehensive review of the literature can be found in Appendix A.2. Here, we propose some extensions to the algorithms in §3.1. Consider the proposed difference LASSO algorithm, the prediction using random forests corresponds to a single imputation. We now propose an approximate "multiply impute, then select" approach to variable selection, which we term the multiple difference LASSO. The algorithm for the multiple difference LASSO proceeds as follows.

1. Fit a random forest regression model for Y on T and $\mathbf{X}$ using the training dataset. We will denote the model as M_0.
2. Based on the model fitted in step 1., simulate Y using the test dataset based on a normal distribution with mean given by the fitted value and variance given by the average mean squaed error of the regression trees fit in the previous step.
3. Repeat step 2 k times to get k sets of potential outcomes. We can equivalently think of k as the number of imputations to perform. Then compute the derived variable $\tilde{Y}$ for each imputed dataset. Here, a small k is recommended, e.g., $k = 3$ [24]. Then, we have three sets of derived variables.
4. Perform a LASSO based on the average value of $\tilde{Y}$ using `glmnet`.

In the above algorithm, we create three imputed datasets and average the value of $\tilde{Y}$. We then perform one single LASSO on the averaged value. Another idea is to replace step 4 by the following:

4.' Perform separate LASSOs on each imputed dataset using `glmnet`.

If we take this alternative approach, we will have three sets of variables selected from the LASSO method. We need to employ some combining rules to determine which variables are eventually selected. As illustrated in [33], there are several ways regarding variable selection for multiple imputed datasets. For example, we may select a covariate if: (1) the covariate is selected in any model; (2) the covariate is selected in at least half of the models; or (3) the covariate is selected in all the models. While (1) is too optimistic and (3) is too conservative, (2) seems a reasonable choice. That is, when we generate three imputed datasets ($k = 3$), we claim that the covariate is selected if it appears in at least two final models. In the simulation study, we term the above two algorithms as "multiple difference LASSO[1]" and "multiple difference LASSO[2]".

3.3. Practical Issues

We now describe some practical issues in the various LASSO algorithms described in this paper. The first deals with the issue of the imputation/prediction model relative to the regression model being used for the LASSO. With respect to the combining rules of [39] used in the multiple difference LASSO algorithms, one underlying assumption is that the different models are conditioned on the same set of data. Given this assumption, it is

necessary that all variables included in the LASSO model must also be present in the model for imputation by random forests.

One other point we wish to make is that we are also assuming sufficient covariate overlap in the $T = 1$ and $T = 0$ subgroups. If this assumption is violated, then we run risk of imputing potential outcomes based solely on model extrapolation. Thus, it is important to explore the data to determine the overlap in covariate distributions.

Another important issue is to compute the standard errors of the estimated regression coefficients for the selected covariates by LASSO and calculate the 95% confidence interval. This will help us check if a selected covariate is significantly related to treatment heterogeneity. The standard errors of LASSO estimators have been discussed in the literature, such as [42], [43] and [44]. Kyung et al. [43] point out that the traditional pairwise bootstrap method does not yield valid standard errors for LASSO. Here, we will employ the modified residual bootstrap method proposed by Chatterjee & Lahiri [44]. The detailed algorithm is presented in Appendix A.2. In Section 4.1, we apply this method to estimate standard errors of the LASSO estimates.

4. Numerical Examples

4.1. Right-heart catheterization study

Our example is from Connors et al. [40]. The question of interest is whether or not the treatment by right heart catherization (RHC) has an effect on 30-day survival (dead/alive at 30 days). The dataset contains information on 5735 patients, 2184 of whom received RHC. Since we focus on average causal effects here, the scientific parameter of interest is a causal risk difference. We consider 21 variables for inclusion in the modelling from the original 75 that are given in the data based on biological plausibility and temporal ordering. Variables are excluded due to missing values or prevalence of less than 20% in the dataset. We tended to focus on demographic variables as well as biological variables as candidate predictors. We also exclude the variable describing the risk prediction made using a model previously developed by the SUPPORT investigators [41] because this variable dominated the variable selection procedures.

We first randomly divide the whole dataset into two parts with equal sample sizes. We use the first half of the data to find covariates that are related to treatment heterogeneity and then use the second half to estimate the treatment effects. Our first set of analyses illustrates the difference LASSO algorithms. From the results shown in Figure 2, we make several observations. First, the effects of all potential variables are relatively small and that there are many predictors that are driving heterogeneity in the average causal effect. Again, the plot is meant to be used to identify predictors for which subgroups should be defined. We have highlighted three variables that appear to be drivers of the heterogeneity: hematocrit, PH, cardiovascular diagnosis. These are the variables whose paths separate from the rest of the variables, especially as the L_1 norm of the estimated coefficients gets larger. We also employ cross-validation to choose the optimal smoothing parameter λ in (11); with the optimal $\lambda = 0.01989$, the selected variables by LASSO are hematocrit, cardiovascular diagnosis and blood pressure. The standard errors and the 95% confidence interval for the

selected covariates based on 10000 bootstrapped samples are displayed in Table 1. The confidence interval indicates that only blood pressure and hematocrit are significantly related to treatment heterogeneity at $\alpha = 0.05$. The bootstrap algorithm also provides 95% confidence intervals for the remaining 18 variables as well. For all the variables that are not selected, the 95% confidence intervals contain the value of zero. To be noticed, one drawback is that the bootstrap method relies on a single value of the smoothing parameter in LASSO. When we perform the follow-up analysis, we will combine the results from the solution path and the cross-validation approach.

Next, we show the results for the multiple difference LASSO algorithms. The procedure is described in Section 3.2. For completeness, we show the plots of the individual difference LASSO algorithms in Figure 3. The colors of the variables are the same across all plots. The multiple LASSO algorithms allow for a qualitative exploration of the variability in predictor selection for the difference LASSO procedure. This is due to the fact that there is tremendous correlation between the variables and that slight perturbations lead to selections of completely different sets of variables in the LASSO algorithm. In addition, one implicit assumption here is that all the variables have been standardized so that selection is being made effectively using the correlation matrix of the predictors. When we performed the averaging across the individual LASSO output, the results were qualitatively quite similar to the results in Figure 2 (data not shown).

After identifying the covariates that are related to treatment heterogeneity, we then perform a follow-up analysis. As mentioned earlier, we are interested in estimating the causal risk difference between the treatment group (RHC) and the control group (no RHC). The outcome variable “dth30” indicates whether the patient died at 30 days and the treatment variable “swang1” indicates the use of RHC. We divide the sample (which is the test data) into 16 strata according to the values of hematocrit, PH, cardiovascular diagnosis and blood pressure. The four variables are chosen based on the LASSO solution path (Figure 2) as well as the selected variables using the best smoothing parameter. While defining the strata, the continuous variable is dichotomized at its median value. In each stratum, we employ an inverse probability weighting (IPW) method to estimate the causal risk difference. The propensity score used in IPW is estimated from a random forests model using all of the available covariates in the study. The random forests model automatically picks important covariates, nonlinear terms and interaction terms for estimating the propensity scores. Table 2 shows the estimated causal risk difference in each stratum. The standard errors are estimated by the sandwich formula using the `survey` package in R. We find that in 2 out of 16 strata, RHC decreases the chance of survival at 30 days, while in the rest of the strata, RHC does not have a significant effect on survival at a significance level of 0.1.

4.2 Simulation Studies

In this section, we conduct simulation studies to compare the proposed methods with the existing methods. Following the data analysis, we generate 20 covariates: $X_1 - X_{10} \sim \mathrm{N}(0,1)$ and $X_{11} - X_{20} \sim$ Bernoulli(0.5). The binary treatment variable T is generated from a logistic regression model with

$$f(\mathbf{X}) = 0.028X_1 - 3.374X_2 - 0.03X_3 + 0.118X_4 - 1.394X_{11} + 0.875X_{12} + 0.9X_{13}$$

and $P(T = 1) = \text{expit}\{f(\mathbf{X})\}$. We consider two cases for the outcome model:

1. Case 1: $Y = 2.455 + 0.37T + 0.1X_1 - 0.154X_2 - 0.152X_{11} - 0.126X_{12} + \varepsilon$
2. Case 2: $Y = 2.455 + 0.37T + 0.1X_1 - 0.154X_2 - 0.152X_{11} - 0.126X_{12} - 0.3T \times X_{11} + \varepsilon$

where $\varepsilon \sim N(0, \sigma^2)$.

For comparison, we employ five different approaches to identify covariates that define subgroups in which the causal effect is heterogeneous: difference LASSO, multiple difference LASSO[1], multiple difference LASSO[2], difference AIC and traditional LASSO. In both versions of multiple difference LASSO, 3 imputed datasets will be generated in each run. The multiple difference LASSO[1] takes the average value of the difference and performs one LASSO. In contrast, the multiple difference LASSO[2] performs three separate LASSOs for different imputed datasets, and a covariate will be selected if it is selected in at least two imputed datasets. In the difference AIC, the Akaike information criterion instead of LASSO is employed for variable selection. Finally, in the traditional LASSO, we regress the observed Y on all the possible main effects and one-way interactions between the treatment and each covariate. Then, we identify the significant main effects and interaction terms. The smoothing parameter is optimally determined by cross-validation, which can be easily realized in `glmnet`. We run the simulation $N = 1000$ times. For the first four methods, we record the number of simulations in which each covariate is selected. For the traditional LASSO, we record the number of simulations in which each interaction term is selected. The simulation results are shown in Tables 3 and 4.

In the first case, none of the covariates have interactions with the treatment. It means the treatment effect is homogeneous, i.e. assumption (9) is true. In this case, the percentage of simulations in which each covariate is selected corresponds to the type I error. We average the results over all the variables. As shown in Table 3, the two different versions of multiple difference LASSO yield the lowest average type I error for different sample sizes, followed by difference LASSO. The type I error of multiple difference LASSO[2] is much smaller than the others. In the proposed algorithm, If we employ AIC instead of LASSO for variable selection, the average type I error increases significantly due to the fact that AIC tends to select larger models. Another interesting observation is that the traditional LASSO seems to yield the overall largest type I error. This is because the regression model we fit in the traditional LASSO method ignores the selection bias. We tried different σ values ($\sigma = 0.1$, 0.3, 0.5) in the simulation and the results are very similar. In the second case, there is a significant interaction between X_{11} and T in the outcome model, which means the treatment effects are different in the subgroups defined by X_{11}. As a result, there are two columns for each sample size and each σ value in Table 4. The first column records the percentage of simulations in which X_{11} is selected into the model and the second column represents the average percentage of simulations in which other covariates are selected. The former corresponds to (1-type II error) for X_{11} and the latter corresponds to the average type I error.

As can be seen, the two multiple difference LASSOs have the smallest average type I error, followed by difference LASSO. The type I error of the traditional LASSO method is much larger than the other three. This is true for different sample sizes and different σ values. On the other hand, when σ = 0.5, the multiple difference LASSOs yield the largest type II errors, followed by difference LASSO. Again, difference AIC has the smallest type II error because AIC tends to select larger models. When σ decreases to 0.1 (signal-to-noise ratio= 3), all five methods are able to identify the covariate that is related to treatment heterogeneity.

In the imputation procedure, we employ random forests to predict the missing potential outcomes. The reason, as we mentioned earlier, is that random forests can automatically perform variable selection. To examine whether alternative imputation methods would affect the performance, we also employ multiple imputation by chained equations (MICE, [34]) as the imputation method. In addition, we impute the missing potential outcomes from all available predictors and perform a LASSO on the difference in the potential outcomes. The results recorded in Table 3 and 4 indicate that MICE leads to the highest type I error among all the approaches we tested. On the other hand, it produces the smallest type II errors for most of the conditions in Case 2. This is probably due to the fact that there are a large number of redundant features in the simulation setup and our specification for MICE (using all available covariates) ignores this important property while performing imputation.

6. Discussion

In this article, we have explored the use of the LASSO for variable selection in estimating causal effects, which can be applied for identifying subgroups in which treatment effects are homogeneous. We found that within the potential outcomes framework, the variable selection is inherently for a multivariate joint distribution so that what becomes important is the functional of the joint distribution that we seek to model. We consider one functional of interest and attendant LASSO procedures, termed the difference LASSO and multiple difference LASSO. For the latter, we developed two different algorithms based on how to combine the imputed datasets and variable selection results. All proposed procedures have an ‘impute, then select’ structure that is reminiscent of algorithms in the missing data literature [35, 36]. A modified bootstrap method [44] has been used to estimate the standard error and calculate the confidence interval. We can then test the significance of the selected covariates that define subgroups in which the treatment effect is homogeneous. In addition, simulation results indicate that for causal inference, the imputation algorithm for ‘filling in’ the missing potential outcomes is as important to the process as the application of the LASSO. For example, we suggest employing random forests to impute the missing potential outcome, but this needs to be more formalized and explored in future work.

Conceptually, we imagine the use of the LASSO in this context as more of an exploratory device rather than a confirmatory device. As alluded to by the simulation study, when the dimension of the covariates is small or moderate compared to the sample size, AIC/BIC can be used instead of LASSO to increase the power. The purpose of our study is to identify covariates that are related to treatment heterogeneity. Using selected covariates, multiple strata can be defined, and different treatment effects can be estimated within each strata.

This will help researchers make more informed decisions in biomedical or epidemiological studies.

Acknowledgments

The authors would like to acknowledge the efforts of two reviewers, whose comments substantially improved the paper. The research of the first author is funded by NIH CA 129102 and NIDA P50 DA010075.

Appendix

A. 1. Model of Tran et al. [21]

Tran et al. [21] consider the following probabilistic model:

$$\boldsymbol{Y} \sim N(\mathbf{X}\theta, \sigma^2 \mathbf{I}_n) \quad (12)$$

$$\theta|\sigma^2 \sim N(\mathbf{m}, \sigma^2 \mathbf{V}) \quad (13)$$

$$\sigma^2 \sim IG(a, s). \quad (14)$$

Note that model (12)–(14) represents a standard hierarchical linear model. Equation (12) specifies the regression model for **Y**; θ denotes the regression coefficients, and σ^2 is the variance. A prior is assumed for θ and σ^2 based on the product of the densities in (13) and (14). The regression coefficients are assumed to have a normal prior distribution, while the variance parameter σ^2 is assumed to have an inverse gamma distribution in (14). Its density is given by

$$f(\sigma^2) = \frac{(a/2)^{(s/2)}}{\Gamma(s/2)} (\sigma^2)^{-s/2-1} \exp(-\frac{a}{2\sigma^2}).$$

Given this model, Tran et al. [21] show that the predictive distribution for a new observation is given by the t-distribution; the relevant parameter definitions can be found in §3.1. of [21]. They then show that minimizing (7) in §2.3. is equivalent to minimizing

$$\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} E[\{y_i^{new} - (\mathbf{x}_i^{new})'\theta\}^2 | \mathbf{x}_i^{new}, \text{data}] \quad (15)$$

subject to the LASSO constraint on θ. Note that the superscript "new" refers to a future observation, so that the expectation is being taken with respect to the predictive distribution. Further algebraic simplification of (15) reveals that its optimization is equivalent to optimization of

$$\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} s^2 w(\mathbf{x}_i^{new}) + \frac{1}{2\sigma^2}\sum_{i=1}^{n} (\mathbf{x}_i^{new}\theta - \mathbf{x}_i^{new}\hat{\theta})^2 \quad (16)$$

with respect to a LASSO constraint on θ. Formulae defining *s* and *w* can be found in [21]. We also note that the 'LASSO constraint on θ' is in fact a weighted constraint on the sum of the L_1 norms of θ; the weights depend on the predictive distribution.

A. 2. Multiple Imputation and Variable Selection

Yang et al. [35] considered the problem of imputation and variable selection for linear regression models. They focused on using a Bayesian formulation inspired by the stochastic search variable selection algorithm of George and McCulloch [37]. We review the model here:

$$Y_i|\beta,\gamma\sim N(\beta_\gamma^{'}\mathbf{X}_i,\sigma^2\mathbf{I}_{\sum_{j=1}^{p}\gamma_j}) \quad (17)$$

$$\beta_j|\gamma_j\sim(1-\gamma_j)N(0,\tau^2)+\gamma_j N(0,c^2\tau^2) \quad (18)$$

$$\sigma^2|\gamma\sim IG(\nu_\gamma/2,\nu_\gamma\lambda_\gamma/2). \quad (19)$$

Model (17)–(19) specifies a probabilistic model for linear regression. The γ_j represent binary latent indicator variables where a value of one indicates that the variable should be included in the model and zero denotes that it should not. Equation (17) specifies the linear regression model given the selected covariates that are placed in the model. Equation (18) models the regression coefficients as a mixture of normals, conditional on whether the variable is selected or not. The former group of variables will have a larger variance as *c* is typically chosen to be much larger than one [37]. To complete the model, one typically assumes the γ_j to be Bernoulli distributed. While George and McCulloch [37] and Yang et al. [35] both develop simulation-based approaches to Bayesian inference in this model, we will instead use an equivalence between the LASSO with a slightly different version of the above model that was developed by Yuan and Lin [38]. This requires replacing (18) by

$$\beta_j|\gamma_j\sim(1-\gamma_j)\delta(0)+\gamma_j DE(0,\tau),$$

where δ(0) is a point mass distribution at zero and *DE*(0, τ) denotes the double-exponential distribution with mean zero and scale parameter τ. Also, we replace the Bernoulli assumption on γ by the following prior for γ:

$$P(\gamma)\propto q^{\sum_{j=1}^{p}\gamma_j}(1-q)^{p-\sum_{j=1}^{p}\gamma_j}\sqrt{det(\mathbf{X}_\gamma'\mathbf{X}_\gamma)}. \quad (20)$$

Based on this variation of the George-McCulloch model, Yuan and Lin [38] show that the model of the posterior distribution can be estimated using the LASSO algorithm.

Within this model framework, Yang et al. [35] considered the case of missing predictor variables and proposed two types of procedures. The first was termed "impute, then select" (ITS). For ITS methods, the analyst first imputes the data and creates several imputed datasets. Then one performs the stochastic variable search algorithm of George and

McCulloch [37] in parallel and applies mutiple imputation combining rules [39] in order to perform variable selection. The other class of methods proposed by [35] was a simutaneously impute and select (SIAS) algorithm. Here, one performs imputation and variable selection within one larger iterative algorithm. While the SIAS method was recommend by Yang et al. [35], they also noted its computational intensiveness. Wood et al. [33] focused on multiple imputation-based procedures and found through simulation studies the strategy of "multiply impute, then select" (MITS) to perform the best. The idea of MITS is to consider several imputed datasets, perform variable selection and to then combine variable selection results across the datasets using the combining rules in [39]. It is obvious that that the MITS is the multiple imputation analog of the ITS approach of Yang et al. [35]. The work in Yang et al. [35] and Wood et al. [33] deal with the issue of missing covariates, whereas we are dealing with missing *Y* values, and in particular, missing potential outcomes.

A. 3. Modified Bootstrap Method for LASSO Estimates

Following Chatterjee & Lahiri (2011) [44], the modified bootstrap approach to estimate the standard error and the confidence interval for LASSO is as follows:

1. Perform LASSO based on $(X_i, \tilde{Y}_i)$, $i = 1, \ldots, n$ and get the estimated coefficient vector $\tilde{\gamma} = (\tilde{\gamma}_1, \tilde{\gamma}_2, \ldots, \tilde{\gamma}_p)'$ in (11). Denote the "best" smoothing parameter selected by cross-validation as $\hat{\lambda}$.
2. Shrink the non-zero components of $\tilde{\gamma}$ to zero if it is smaller than $a_n = cn^{-\delta}$ and denote the modified estimate of γ as $\tilde{\gamma}$. For example, in the data application, we let $\delta = 1/3$ and $c = 1$.
3. Calculate $e_i = \tilde{Y}_i - \hat{\beta}_0 - r'\mathbf{X}_i$, $i = 1, \ldots, n$ and centralize e_i so that $\sum_{i=1}^{n} e_i = 0$.
4. Take a bootstrap sample of e_i with replacement and denote as $(e_1^*, e_2^*, \ldots, e_n^*)$.
5. Calculate $\tilde{Y}^* = \hat{\beta}_0 + \hat{r}'\mathbf{X}_i + e_i^*$ $i = 1, \ldots, n$. Perform LASSO based on $(X_i, \tilde{Y}_i^*)$ with smoothing parameter $\hat{\lambda}$ and get the LASSO estimates as $\hat{r}^*$.
6. Repeat step 4 and 5 N times and denote the estimated coefficient vectors as $r^{*1}, r^{*2}, \ldots, r^{*N}$.

The standard error for the *j*th component of $\hat{\gamma}$ is calculated as

$$SE = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\gamma}_j^{*i} - \bar{\gamma}_j^*)^2}$$

where $\bar{\gamma}_j^* = \sum_{i=1}^{N} \hat{\gamma}_j^{*i}/N$ and the $(1 - \alpha)$ confidence interval for γ_j is calculated as

$$(\tilde{\gamma}_j + \hat{\gamma}_j - \hat{\gamma}_{j,(1-\alpha/2)}^*, \tilde{\gamma}_j + \hat{\gamma}_j - \hat{\gamma}_{j,\alpha/2}^*)$$

where $\hat{\gamma}_{j,\alpha}^*$ is the αth percentile of $\hat{\gamma}_j^{*1}, \hat{\gamma}_j^{*2}, \ldots, \hat{\gamma}_j^{*N}$.

References

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66:688–701.
2. Holland P. Statistics and causal inference (with discussion). Journal of the American Statistical Association. 1986; 81:945–970.
3. Witteman JC, D'Agostino RB, Stijnen T, Kannel WB, Cobb JC, de Ridder MA, Hofman A, Robins JM. G-estimation of causal effects: isolated systolic hypertension and cardiovascular death in the Framingham Heart Study. American Journal of Epidemiology. 1998; 148:390–401. [PubMed: 9717884]
4. Stuart EA. Matching methods for causal inference: a review and a look forward. Statistical Science. 2010; 25:1–21. [PubMed: 20871802]
5. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–972. [PubMed: 16401269]
6. Zhu Y, Ghosh D, Mitra N, Mukherjee B. A data-adaptive strategy for inverse weighted estimation of causal effects. Health Services Outcomes and Research Methodology. 2014; 14:69–91.
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70:41–55.
8. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine. 2004; 23:2937–2960. [PubMed: 15351954]
9. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B. 1996; 58:267–288.
10. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.
11. Hirano K, Imbens G. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Services and Outcomes Research Methodology. 2001; 2:259–278.
12. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. American Journal of Epidemiology. 2006; 163:149–56.
13. Crainiceanu C, Dominici F, Parmigiani G. Adjustment uncertainty in effect estimation. Biometrika. 2008; 95:635–651.
14. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. Statistical Methods in Medical Research. 2014; 21:7–30. [PubMed: 21075803]
15. Wang C, Parmigiani G, Dominici F. Bayesian effect estimation accounting for adjustment uncertainty (with discussion). Biometrics. 2012; 68:661–676. [PubMed: 22364439]
16. Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics and Data Analysis. 2006; 50:475–498.
17. Van der Laan, MJ.; Rose, S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer; 2011.
18. Pearl, J. Causality: Models, Reasoning and Inference. Cambridge University Press; 2009.
19. Bühlmann P, Kalisch M, Maathuis MH. Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. Biometrika. 2010; 97:261–278.
20. Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association. 1996; 91:473–489.
21. Tran MN, Nott DJ, Leng C. The predictive Lasso. Statistics and Computing. 2012; 22:1069–1084.
22. Robins J. Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics, Theory and Methods. 1994; 23:2379–2412.
23. Robins JM, Hernán M, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–560. [PubMed: 10955408]
24. Little, RJA.; Rubin, DB. The Statistical Analysis of Missing Data. 2nd edition. New York: Wiley and Sons; 2002.

25. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003; 14:680–686. [PubMed: 14569183]
26. Geisser S. The predictive sample reuse method with applications. Journal of the American Statistical Association. 1975; 70:320–328.
27. Van Buuren, S. Flexible Imputation of Missing Data. Boca Raton, FL: Chapman & Hall/CRC Press; 2012.
28. Raghunathan TE, Lepkowski JM, Van Hoewyk JV, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 2001; 27:85–95.
29. Biau G, Devroye L, Lugosi G. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research. 2008; 9:2015–2033.
30. Breiman L. Random forests. Machine Learning. 2001; 45:5–32.
31. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33
32. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics. 1985; 41:361–372. [PubMed: 4027319]
33. Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. Biometrics. 2005; 61:498–506. [PubMed: 16011697]
34. Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45
35. Wood AM, White IR, Royston P. How should variable selection be performed with mutiply imputed data? Statistics in Medicine. 2008; 27:3227–3246. [PubMed: 18203127]
36. Chen Q, Wang S. Variable selection for multiply-imputed data with application to dioxin exposure study. Statistics in Medicine. 2013; 32:3646–3659. [PubMed: 23526243]
37. George E, McCulloch RE. Variable selection with Gibbs sampling. Journal of the American Statistical Association. 1993; 88:881–889.
38. Yuan M, Lin Y. Efficient Empirical Bayes variable selection and estimation in linear models. Journal of the American Statistical Association. 2005; 100:1215–1225.
39. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley; 1987.
40. Connors AF Jr, Speroff T, Dawson NV, Thomas C, Harrell FE Jr, Wagner D, Desbiens N, Goldman L, Wu AW, Califf RM, Fulkerson WJ Jr, Vidaillet H, Broste S, Bellamy P, Lynn J, Knaus WA. The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators. Journal of the American Medical Association. 1996; 276:889–897. [PubMed: 8782638]
41. Knaus WA, Harrell FE Jr, Lynn J, Goldman L, Phillips RS, Connors AF Jr, Dawson NV, Fulkerson WJ Jr, Califf RM, Desbiens N, Layde P, Oye RK, Bellamy PE, Hakim RB, Wagner DP. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. Annals of Internal Medicine. 1995; 122:191–203. [PubMed: 7810938]
42. Knight K, Fu W. Asymptotics for lasso-type estimators. Annals of statistics. 2000; 28:1356–1378.
43. Kyung M, Gill J, Ghosh M, Casella G. Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis. 2010; 5:369–411.
44. Chatterjee A, Lahiri SN. Bootstrapping lasso estimators. Journal of the American Statistical Association. 2011; 106:608–625.

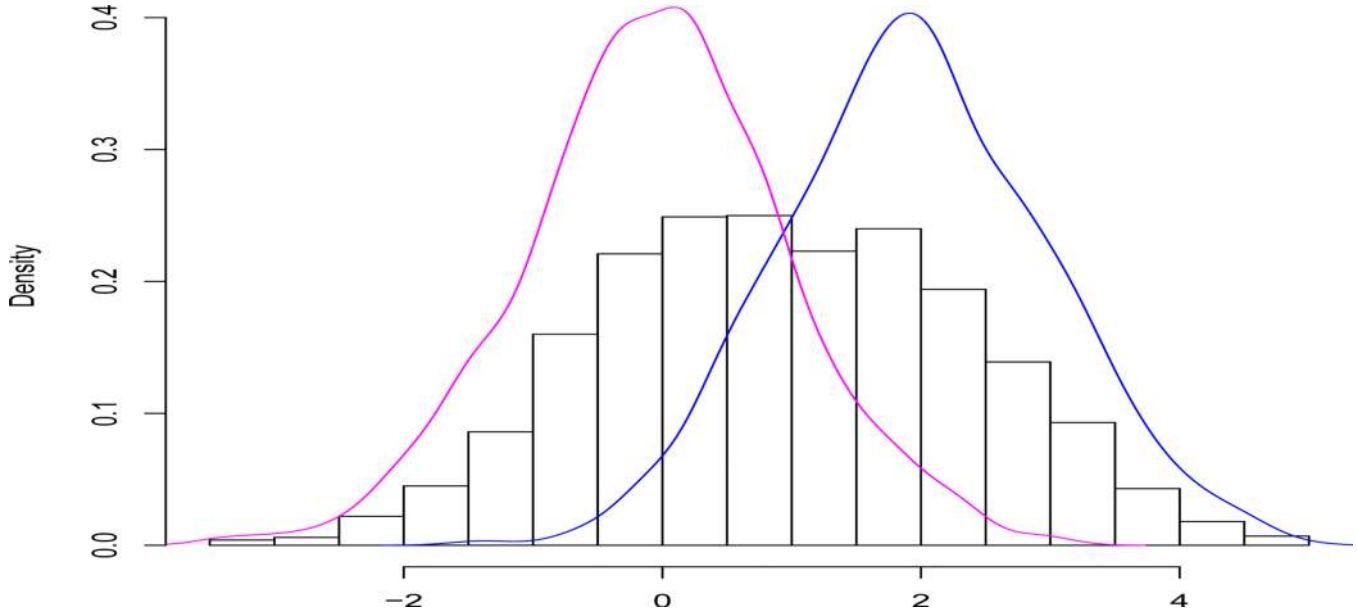

Figure 1.

Distribution of covariate X for treatment and control groups. The blue line denotes the kernel density estimation for X in the $T = 1$ group, while the magenta line represents the kernel density estimate for X in the $T = 0$ group. The bars represent the histogram of X regardless of the treatment groups.

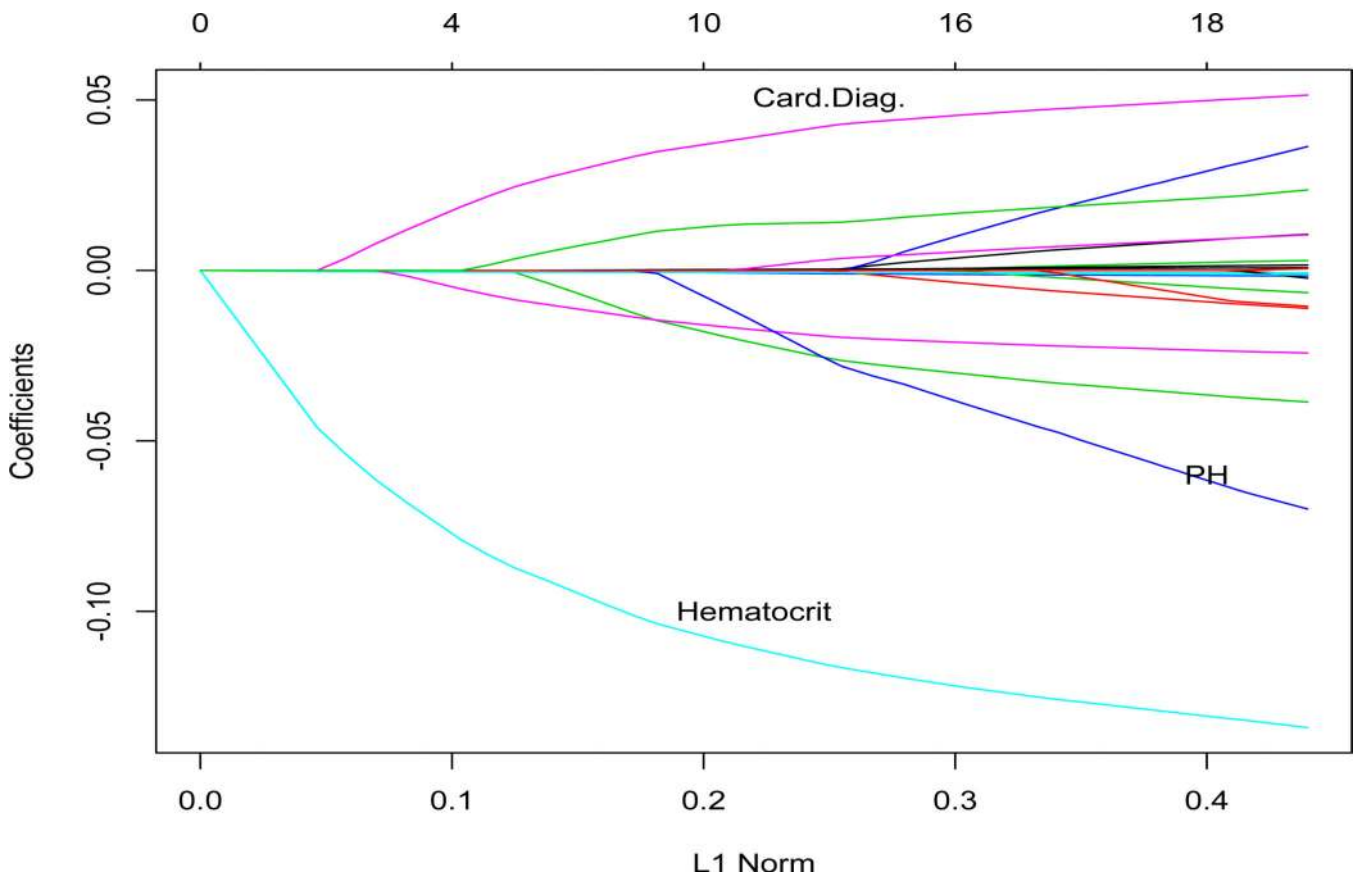

Figure 2.

Output for the difference LASSO algorithm. The regularized solution paths of the regression coefficients for all of the variables as a function of the L_1 norm of the estimated coefficients are plotted.

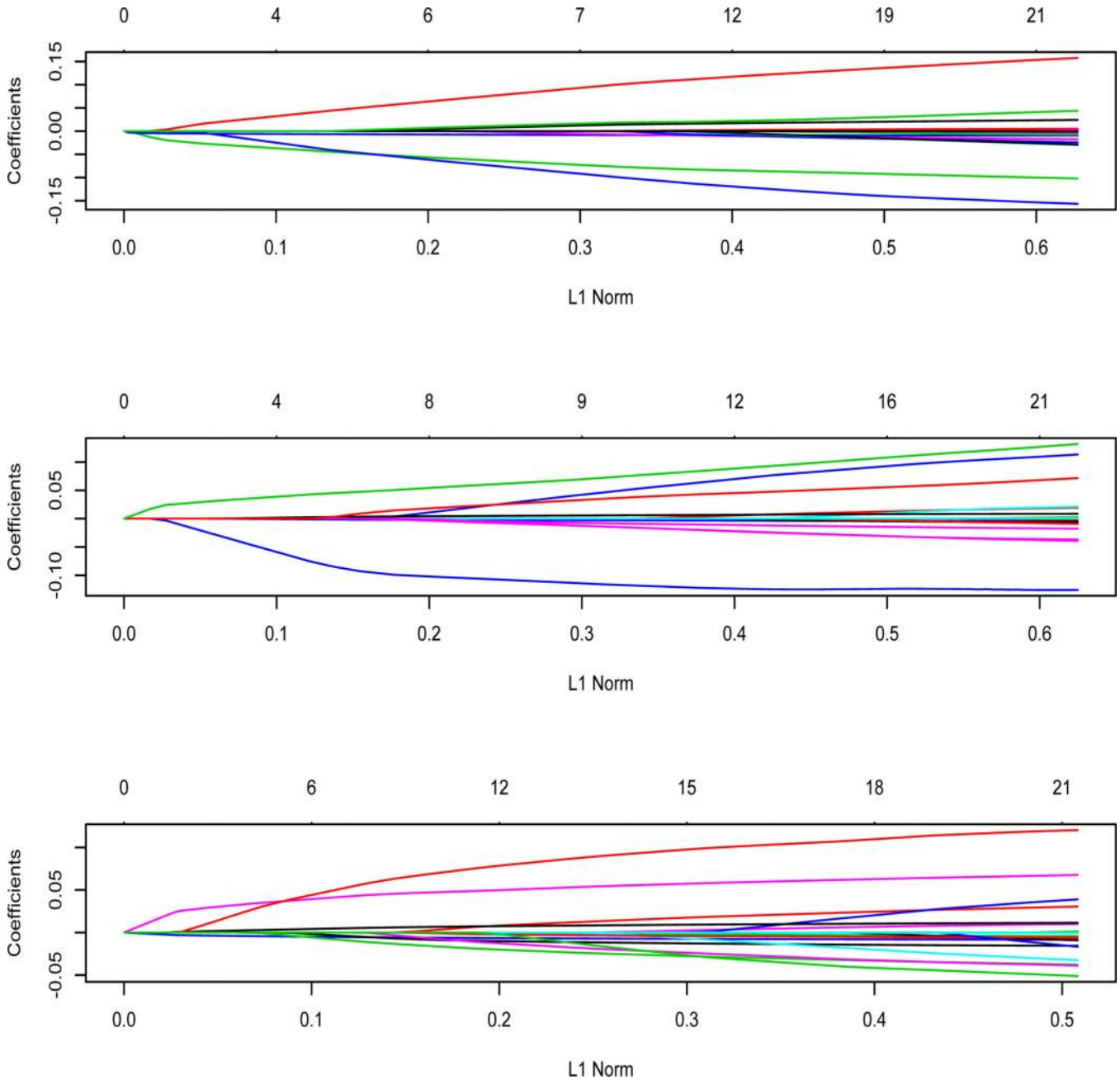

Figure 3.

Output for the difference LASSO algorithm. The regularized solution paths of the regression coefficients for all of the variables as a function of the L_1 norm of the estimated coefficients are plotted.

Table 1

Estimated Coefficients by LASSO and the Bootstrap Standard Errors

Covariate	Estimated Coefficient	Bootstrap SE	95% Confidence Interval
bp	-3.86×10^{-4}	1.86×10^{-4}	$(-7.73 \times 10^{-4}, -1.34 \times 10^{-4})$
hema	-6.13×10^{-2}	2.78×10^{-2}	$(-1.23 \times 10^{-1}, 2.63 \times 10^{-2})$
card	7.84×10^{-3}	7.56×10^{-3}	$(-7.33 \times 10^{-3}, 2.53 \times 10^{-2})$

Table 2

Causal Risk Difference Estimates in Different Strata

Strata	bp< 78.2, card=Yes	bp< 78.2, card=No	bp ≥78.2, card=Yes	bp ≥78.2, card=No
hema< 30, ph< 7.4				
Estimate (s.e.)	0.160(0.083)	0.089(0.053)	−0.159(0.127)	−0.003 (0.102)
p-value	0.056	0.093	0.217	0.973
hema< 30, ph ≥7.4				
Estimate (s.e.)	0.047 (0.088)	−0.052(0.058)	0.062(0.136)	0.003(0.081)
p-value	0.597	0.368	0.649	0.971
hema ≥30, ph< 7.4				
Estimate (s.e.)	0.018(0.069)	−0.001(0.065)	0.139(0.108)	0.094(0.162)
p-value	0.792	0.991	0.200	0.560
hema ≥30, ph ≥7.4				
Estimate (s.e.)	0.042(0.074)	0.025(0.077)	−0.152(0.092)	−0.051 (0.147)
p-value	0.574	0.740	0.104	0.730

Table 3

Simulation Results for Case 1 (%)

Method	***n* = 200**	***n* = 500**	***n* = 1000**
Difference LASSO	9.8	12.3	13.5
Multiple Difference LASSO[1]	9.6	10.3	12.2
Mutilple Difference LASSO[2]	4.5	4.6	6.0
Difference AIC	26.0	19.8	19.4
Traditional LASSO	20.3	22.2	29.4
MICE	41.2	40.6	42.1

Table 4

Simulation Results for Case 2 (%)

Method	***n* = 200**		**n = 500**		***n* = 1000**	
σ = 0.1	X_{11}	Others	X_{11}	Others	X_{11}	Others
Difference LASSO	100.0	18.2	100.0	19.1	100.0	20.5
Multiple Difference LASSO[1]	99.7	18.0	100.0	18.9	100.0	20.5
Multiple Difference LASSO[2]	99.0	10.7	100.0	12.7	100.0	13.7
Difference AIC	100.0	21.2	100.0	20.2	100.0	22.1
Traditional LASSO	100.0	37.8	100.0	37.9	100.0	38.3
MICE	100.0	46.6	100.0	47.5	100.0	47.9
σ = 0.3	X_{11}	Others	X_{11}	Others	X_{11}	Others
Difference LASSO	72.1	15.1	99.3	18.8	100.0	19.2
Multiple Difference LASSO[1]	60.7	14.0	97.7	17.8	100.0	19.9
Multiple Difference LASSO[2]	51.6	6.4	94.6	9.4	100.0	11.9
Difference AIC	87.5	20.5	99.9	19.9	100.0	20.7
Traditional LASSO	69.9	33.7	98.2	37.9	100.0	37.1
MICE	90.4	46.8	99.2	47.9	100.0	51.4
σ = 0.5	X_{11}	Others	X_{11}	Others	X_{11}	Others
Difference LASSO	34.1	10.8	76.2	15.5	98.4	19.0
Multiple Difference LASSO[1]	32.0	10.3	68.5	14.2	94.3	17.4
Multiple Difference LASSO[2]	21.6	4.9	53.5	7.0	90.3	10.5
Difference AIC	55.3	20.1	88.8	18.6	99.7	18.2
Traditional LASSO	34.5	26.2	69.7	32.6	94.5	36.5
MICE	70.5	45.8	90.0	46.7	97.7	49.5