2013

Penalized regression procedures for variable selection in the potential outcomes framework

Debashis Ghosh, *Penn State University*
Yeying Zhu, *University of Waterloo*
Donna L Coffman, *Penn State University*

Penalized regression procedures for variable selection in the potential outcomes framework

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

[1]Department of Statistics and Public Health Sciences, Penn State University

University Park, PA, 16802, USA

[2]Department of Statistics and Actuarial Sciences, University of Waterloo

[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, 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 shared LASSO and difference LASSO algorithm are defined, along with their 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; Qualitative interaction; 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 are distributed the same among different treatment groups. Consequently, the treatment is the only factor that may cause differences in the outcome. However, in an observational study, where the treatment assignment is not controlled by a randomization scheme, 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 model for causal effects is the potential outcomes model (Neyman, 1923; Rubin, 1974; Holland 1986). This framework formulates counterfactual random variables that represent the outcome variable under the hypothetical treatments of interest. 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 estimation of causal effects. One commonly used approach is that of propensity scores (Rosenbaum and Rubin, 1983). This quantity is defined as the probability of the treatment given covariates. Based on the estimated approaches, a variety of modelling strategies can be used to estimate causal effects; a good summary of them can be found in Lunceford and Davidian (2004).

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., Tibshirani, 1996; Fan and Li, 2001). Many approaches for variable selection in causal inference have been described recently. A proposal from Hirano and Imbens (2001) 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 Brookhart et al. (2006). Model averaging approaches for causal effects have been advocated by several authors (Crainiceanu et al., 2008; Vansteelandt et al., 2012; Wang et al., 2012). 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 (2006); this is also related to the general targeted learning framework that has been summarized in the recent monograph by van der Laan and Rose (2011). For the causal graphical modelling framework of Pearl (2009), which is based on acyclic directed graphs, Bühlmann et al. (2010) 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 of Tibshirani (1996) and makes use of the predictive nature of the causal inference problem. 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 (Tran et al., 2012) 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 apply off-the-shelf variable selection procedures to attempt to perform proper causal inference. The concept of predictive LASSO (Tran et al., 2012) 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 two novel LASSO innovations, the shared LASSO and the difference LASSO. Extensions to these procedures, 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 $Y_i = Y_i(T_i)$ $(i = 1, \ldots, n)$, which implies that $Y(0)$ and $Y(1)$ can not be observed simultaneously, i.e. one of them is missing. The relationship between Y and $\{Y(0), Y(1)\}$ can be summarized as

$$Y = Y(1) \times T + Y(0) \times (1 - T). \tag{1}$$

Two parameters of interest are the average causal effect:

$$ACE = E[Y(1) - Y(0)], \tag{2}$$

and the average causal effect among the treated:

$$ACET = E[Y(1) - Y(0)|T = 1]. \tag{3}$$

In a randomized study, the treatment assignment is completely determined by a randomization scheme, which is not dependent on the potential outcomes. Therefore, we have T is independent of $\{Y(0), Y(1)\}$. Consequently, an unbiased estimator for both ACE and ACET is given by

$$\widehat{ACE} = \widehat{ACET} = \frac{\sum_i Y_i T_i}{\sum_i T_i} - \frac{\sum_i Y_i(1 - T_i)}{\sum_i (1 - T_i)}. \tag{4}$$

In a 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 (1983) that allows for proper causal inference in non-randomized observational studies is called strongly ignorable treatment assignment:

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

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 (1983) further proposed the concept of propensity scores, which is defined as the probability of receiving the treatment given the covariates:

$$e(\mathbf{X}) \equiv P(T = 1|\mathbf{X}) \tag{5}$$

Rosenbaum and Rubin (1983) show that if (5) holds, the following property is also true:

$$T \perp \{Y(0), Y(1)\}|e(\mathbf{X}). \tag{6}$$

Since $e(\mathbf{X})$ is a scalar quantity, Rosenbaum and Rubin (1983) 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 wish to 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 to the former 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 Lunceford and Davidian (2004).

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 is 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 (Rosenbaum and Rubin,

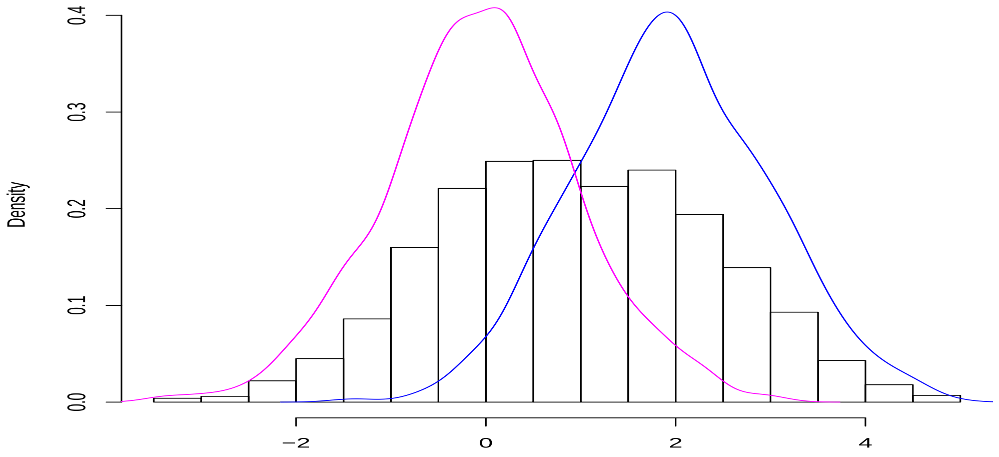

Figure 1: Distribution of covariate X for treatment and control groups. The blue line denotes the kernel density estimation for X in the $T = 1$ group, while the magenta line represents the kernel density estimate for X in the $T = 0$ group.

1983). 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 corresponding 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, these models 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 (Robins, 1994) or marginal structural models (Robins et al., 2000). These differences are at the level of model *specification* and in defining the 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 as described above. 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 (2002). The causal assumptions described in §2.1. and in particular, the assumption (5) correspond to missingness at random (Little and Rubin, 2002), where the missingness mechanism depends on $\mathbf{X}$. This leads to the following conceptually 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

$$\widehat{ACE} = n^{-1}\left[\sum_{i:R_i=1}(Y_i - Y(0)_i^*) + \sum_{i:R_i=0}(Y(1)_i^* - Y_i)\right], \tag{7}$$

where the asterisks in (7) 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 with marginal structural models have been treated in a very simple case by Sato and Maruyama (2003). The reliance on imputation for construction of counterfactuals is intimately tied to the use of the predictive distribution (Geisser, 1975), which will inform our modelling strategy in the next section.

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

As discussed in Lunceford and Davidian (2004), 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 (Tibshirani, 1996), which minimizes the residual sum of squares

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

where $(\beta_0, \beta_1, \gamma)$ are regression coefficients and $t \geq 0$. Note that the constraint in (8) is on the L_1 norms 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 (8) 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 (8). By predictive criterion functions, what we refer to are functions of predicted/imputed values of the response. Our general approach is solve the following optimization problem (Tran et al., 2012):

$$\min_{\alpha} \sum_{i=1}^{n} D_i(M_{full}, M_\alpha) + \lambda \sum_{j=1}^{p} |\alpha_j|, \tag{9}$$

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

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

where f and g represent densities of the data. While the optimization problem in (9) is written down in a general form, what Tran et al. (2012) show is that in a linear model case, it corresponds to solving a weighted lasso problem of the form minimizing

$$\sum_{i=1}^{n} (\hat{\beta}_0 - \hat{\beta}_1 T_i - \hat{\gamma}' \mathbf{X}_i - \beta_0 - \beta_1 T_i - \gamma' \mathbf{X}_i)^2 \text{ s.t. } |\beta_0| + |\beta_1| + \sum_{j=1}^{p} |\gamma_j| \leq t, \tag{10}$$

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 (10) 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. (2012) prove their result using a canonical hierarchical normal model with conjugate priors, and in fact the objective function in (10) arises from the posterior predictive distribution in the model. Further details on the model used by Tran et al. (2012) are given in the Appendix. Another way to reinterpret their algorithm is as follows:

1. Using a training dataset, fit a linear regression of Y on T and $\mathbf{X}$.

2. Based on the model fitted in step 1., compute predicted/fitted values of Y using the test dataset.

3. Perform a LASSO of the fitted values from the previous step on T and $\mathbf{X}$ 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 in which the linear regression is substituted with any arbitrary imputation algorithm. Examples of imputation methods include multiply imputed chained equations (van Buuren, 2012) and IVEWARE (Raghunathan et al., 2001).

3. Proposed methods

3.1. Adaptation to causal inference

We now seek to apply the predictive LASSO for the estimation of causal effects. The first two steps of our algorithm are as follows:

1. Fit a regression model for Y on $\mathbf{X}$ for individuals with $T = 0$ and $T = 1$ separately. This will yield two prediction models, one for the treated group ($T = 1$), and one for the control group ($T = 0$). We will denote these as models M_0 and M_1.

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 M_1 to impute $Y(1)$ for subjects with $T = 0$.

This is very similar to the first two steps in the algorithmic interpretation of the predictive LASSO. One difference is that we fit two separate models here, instead of just one model as in the previous section. Here, we will use random forests (Breiman, 2001) as our prediction algorithm. It belongs to the category of so-called ensemble methods: instead of generating one classification 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 votes for one of the classes and the prediction can be made by the majority votes among the trees. In the first stage of causal inference, if we apply random forests algorithm, the propensity score could be estimated as the proportion of trees that vote for class 1. Biau et al. (2008) proved the consistency of random forests estimator in terms of predicting the class label. In that paper, they also commented that random forests are among the most accurate general-purpose classifiers available.

The application of the LASSO method now requires some care. This is because we fundamentally have a multivariate response for each individual. In particular, we observe 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. Thus, its use is intimately tied to the type of variable selection we wish to perform. The procedures we propose here are quite simple in nature. The first LASSO method we describe is termed the shared LASSO. It creates a univariate response variable for each individual, $\tilde{Y}$, which is the average of the observed and imputed potential outcome, and performs a LASSO of $\tilde{Y}$ on the covariates. Because we are averaging the two variables, the goal of this LASSO procedure is to identify variables that are important to both potential outcomes. Formally, we are solving the following optimization problem:

$$\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{\beta}_0 - \tilde{\gamma}'\mathbf{X}_i)^2 \text{ s.t. } \tilde{\beta}_0 + \sum_{j=1}^{p} |\tilde{\gamma}_j| \leq \tilde{t}, \tag{11}$$

where $(\tilde{\beta}_0, \tilde{\gamma})$ are unknown regression coefficients, and $\tilde{t} \geq 0$ is a smoothing parameter. There exists attendant software in R, glmnet (Friedman et al., 2010), that finds the entire regularization path for the LASSO solution in terms of $\tilde{t}$, to an optimization problem (11), and we use that here.

Another feature of the multivariate response/variable selection being considered here is that causal treatment heterogeneity, as defined by measured covariates, can be identified. We go back to the definition of average causal effect given in (2). This estimand, along with the estimator (4), corresponds to the following model:

$$Y_i(1) - Y_i(0) = \tau + \epsilon_i \tag{12}$$

where $\epsilon_1, \ldots, \epsilon_n$ is a random sample from a distribution with mean zero and variance σ^2. The model (12) 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 (12) is invalid. One can then use the LASSO to identify candidate variables that define subgroups in which the causal effect is heterogeneous. The algorithm is a slight modification of what was proposed earlier in this section. The only difference is that instead of creating $\tilde{Y}$, one would use the derived variable Y^*, defined to be $Y(1) - Y(0)$. We term this approach the difference LASSO.

Intuitively, what drives the difference LASSO procedure in this setting would be the differences in the response that are associated with measured covariates. The selected variables represent the covariates that have what is known as qualitative interaction (Gail and Simon, 1985) with treatment. These are interactions in which the treatment effects have opposite directions in different subgroups defined by the predictor.

3.2. Imputation and variable selection

The algorithms 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 (Yang et al., 2005; Wood et al., 2008; Chen and Wang, 2013). Here, we review that literature and propose some extensions to the algorithms of §4.1.

Yang et al. (2005) considered the problem of imputation and variable selection for linear regresion models. They focused on using a Bayesian formulation inspired by the stochastic search variable selection algorithm of George and McCulloch (1993). We review the model here:

$$Y_i|\beta, \gamma \sim N(\mathbf{X}'_{i,\gamma}\beta_\gamma, \sigma^2\mathbf{I}_{\sum_{j=1}^p \gamma_j}) \quad (13)$$

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau^2) + \gamma_j N(0, c^2\tau^2) \quad (14)$$

$$\sigma^2|\gamma \sim IG(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2). \quad (15)$$

Model (13)-(15) 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 (13) specifies the linear regression model given the selected covariates that are placed in the model. Equation (14) 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 (George and McCulloch, 1993). To complete the model, one typically assumes the γ_j to be Bernoulli distributed. While George and McCulloch (1993) and Yang et al. (2005) 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 (2005). This requires replacing (14) by

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

where $\delta(0)$ is a point mass distribution at zero and $DE(0,\tau)$ 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:

$$P(\gamma) \propto q^{\sum_{j=1}^{p}\gamma_j}(1-q)^{p-\sum_{j=1}^{p}\gamma_j}\sqrt{det(\mathbf{X}'_\gamma \mathbf{X}_\gamma)}. \tag{16}$$

Based on this variation of the George-McCulloch model, Yuan and Lin (2005) show that the mode of the posterior distribution can be estimated using the LASSO algorithm.

Within this model framework, Yang et al. (2005) 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 severay imputed datasets. Then one performs the stochastic variable search algorithm of George and McCulloch (1993) in parallel and applies mutiple imputation combining rules (Rubin, 1987) in order to perform variable selection. The other class of methods proposed by Yang et al. (2005) 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. (2005), they also noted its computational intensiveness. Wood et al. (2008) 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 Rubin (1987). It is obvious that that the MITS is the multiple imputation analog of the ITS approach of Yang et al. (2005). The work in Yang et al. (2005) and Wood et al. (2008) deal with the issue of missing covariates, whereas we are dealing with missing Y values, and in particular, missing potential outcomes.

Consider the algorithms in the previous section. The prediction using random forests corresponds to a single imputation. We now propose an approximate MITS approach to variable selection, which we term the multiple shared LASSO. The algorithm for the mutliple shared LASSO proceeds as follows.

1. Fit a regression model for Y on $\mathbf{X}$ for individuals with $T = 0$ and $T = 1$ separately. This will yield two prediction models, one for the treated group ($T = 1$), and one for the control group ($T = 0$). We will denote these as models M_0 and M_1.

2. Based on the models fitted in step 1., simulate Y based on a normal distribution 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 M_1 to impute $Y(1)$ for subjects with $T = 0$. We will describe how to do this for the right-heart catheterizatio data in §4.

3. Perform a LASSO based on the average of the fitted values using glmnet. There should be three LASSOs performed here.

4. Average the regularization paths from the previous step.

Similarly, one could compute a multiple difference LASSO to identify causal interactions by modifying step 3 and creating a derived variable that is the average of the difference between the potential outcomes.

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 Rubin (1987) used in the multiple shared and 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 prediction model (estimated by random forests) must also be present in the model fit using LASSO.

This article focuses on the use of LASSO for variable/model selection in causal inference. An important issue in fitting any regression model becomes performing inference about regression coefficients. Standard methods of inference for regression models cannot be used here because of the presence of the prediction model being used to impute Y. One suggestion that should be explored further is to use standard multiple imputation rules. Assume we have performed the imputation K times. If $\hat{\theta}^k$ and $\hat{V}^k$ generically denote the regression coefficient estimates and corresponding variance-covariance matrix for the kth imputed dataset, then the estimate of the regression coefficients is given by $K^{-1}\sum_{k=1}^{K}\hat{\theta}^k$ with an attendant variance-covariance matrix given by

$$\mathbf{W} + (1 + K^{-1})\mathbf{B},$$

where $\mathbf{W} \equiv K^{-1}\sum_{k=1}^{K}\hat{V}^k$ is a within-imputation variance and

$$\mathbf{B} \equiv K^{-1}\sum_{k=1}^{K}(\hat{\theta}^k - \bar{\theta})(\hat{\theta}^k - \bar{\theta})^T$$

denotes a between-variance estimator. This is a standard formula for imputation estimators and one that we use for practical computational convenience. However, it should also be noted that if the same dataset is used for the entire algorithm, there might be overfitting issues.

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.

4. Numerical Examples

4.1. Right-heart catheterization study

Our first example is from Connors et al. (1996). The question of interest is whether or not the treatment, 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. Variables are excluded due to missing values or prevalence of less than 20% in the dataset. We also 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 (Knaus et al., 1995) because this variable dominated the variable selection procedures.

Our first set of analyses illustrate the shared and difference LASSO algorithms. The results are given in Figure 2. From Figure 2, we make several observations. First, the scales of the two

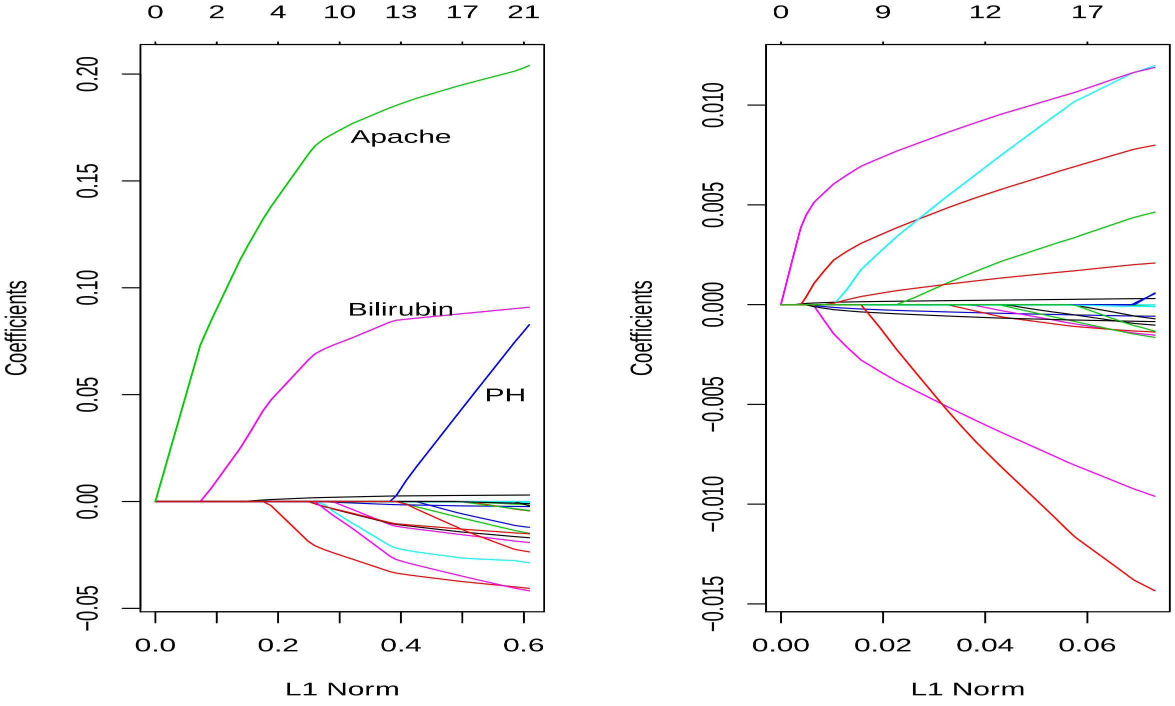

Figure 2: Output for the shared (left) and difference (right) LASSO algorithms. What is plotted are the regularized solution paths of the regression coefficients for all of the variables as a function of the smoothing parameter or equivalently the LASSO constraint parameter.

pictures are quite different. In particular, the one on the right for the difference LASSO is on a much smaller scale than the one on the left. However, what is driving the scale for the shared LASSO algorithm are three predictors that have been marked in the text, Apache score, bilirubin and pH. The former variable is a measure of functional status and has typically to be a strong predictor of mortality. The second variable is a measure of liver function and is also quite reflective of the health status of the subject. What we can infer from the right-hand plot is that there are many predictors that are driving heterogeneity in the average causal effect. Again, the right-hand plot is meant to be used to identify predictors for which subgroups should be defined. Interestingly, two of the predictors that can be used to identify subgroups are bilirubin and hematocrit. Thus, bilirubin appears in both plots. However, we stress that the goals of the two figures are different. The left-hand figure is attempting to determine what variables are important in the generation of the potential outcomes, while the right-hand figure is attempting to determine what predictors should be used to define subgroups for the causal parameter of interest.

Next, we show the results for the multiple shared and multiple difference LASSO algorithms. The procedure proceeded as follows for the multiple shared LASSO:

1. Fit random forests regression for the dependent variable the predictors in the RHC and no-RHC groups.

2. Based on the models fitted in step 1., simulate Y based on a normal distribution with mean given by the fitted value and variance given by the average mean squared error of the regression trees fit in the previous step.

3. Repeat step 2 3 times to get three sets of potential outcomes. The compute the derived variables $\tilde{Y}$ and Y^* for each imputed dataset. There should now be three sets of derived variables.

4. Perform a LASSO based on the average of the fitted values using glmnet.

5. Repeat steps 1-3 and average the regularization paths.

For completeness, we show the plots of the individual shared and difference LASSO algorithms in

Figure 3. The colors of the variables are the same across all plots. What the multiple LASSO algorithms also allow for is a qualitative exploration of the variability in predictor selection for both the shared and difference LASSO procedures. 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).

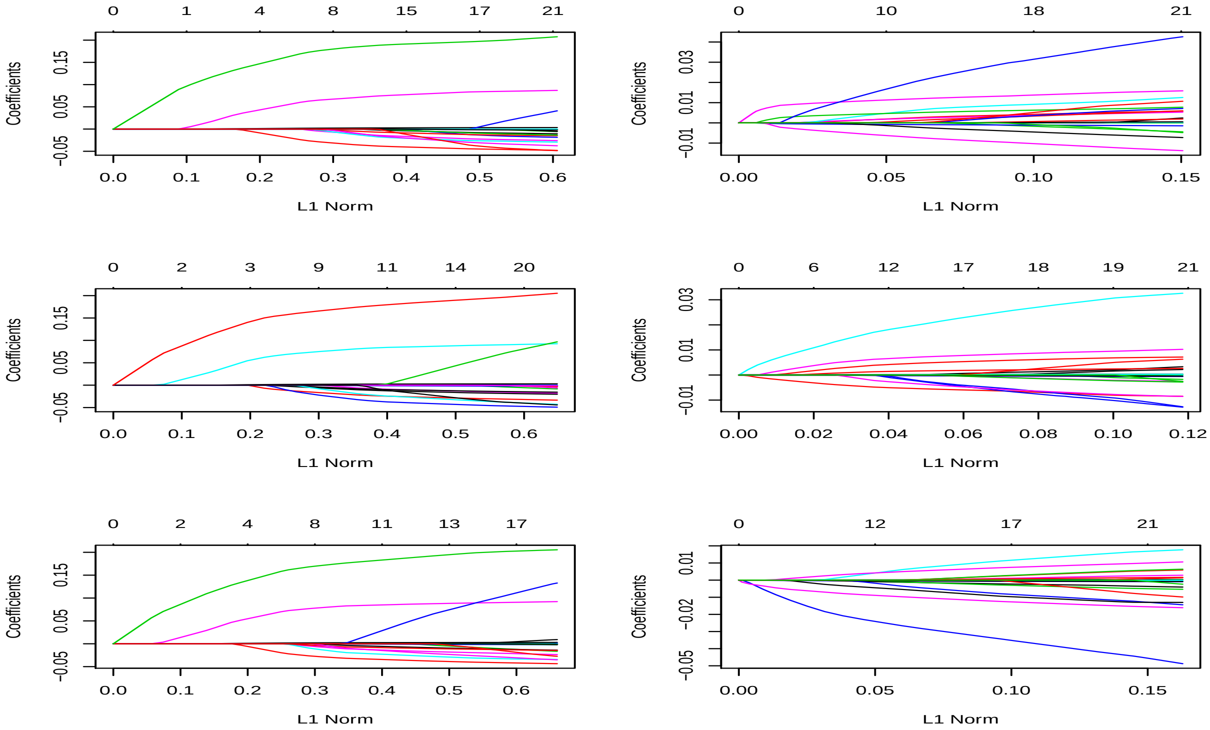

Figure 3: Output for the shared (left-hand and difference (right) LASSO algorithms. What is plotted are the regularized solution paths of the regression coefficients for all of the variables as a function of the smoothing parameter or equivalently the LASSO constraint parameter.

5. Discussion

In this article, we have explored the use of the LASSO for causal effects. 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 propose two functionals of interest and attendant LASSO procedures, termed the shared and difference LASSO. Both procedures have an 'impute, then select' structure that is reminiscent of algorithms in the missing data literature (Yang et al., 2005; Chen and Wang, 2013).

The algorithms proposed in this paper are quite simple conceptually. They exploit ideas from the missing data literature, and this needs to explored further. For example, the question of standard error estimation was brought up in §3.3., but the proposal needs to be rigorously evaluated both theoretically as well as computationally.

Appendix

Model of Tran et al. (2012)

Tran et al. (2012) consider the following probabilistic model:

$$\mathbf{Y} \sim N(\mathbf{X}\theta, \sigma^2\mathbf{I}_n) \tag{17}$$

$$\theta|\sigma^2 \sim N(\mathbf{m}, \sigma^2\mathbf{V}) \tag{18}$$

$$\sigma^2 \sim IG(a, s). \tag{19}$$

Note that model (17)-(19) represents a standard hierarchical linear model. Equation (17) specifies the regression model for $\mathbf{Y}$; 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 (18) and (19). 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 (19). 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. (2012) 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 Tran et al. (2012). They then show that minimizing (9) in §2.3. is equivalent to minimizing

$$\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} E\left[\{y_i^{new} - (\mathbf{x}_i^{new})'\theta\}^2|\mathbf{x}_i^{new}, \text{data}\right] \tag{20}$$

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 (20) reveals that its optimization is equivalent to optimization of

$$\frac{n}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} s^2 w(\mathbf{x}_i^{new}) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\mathbf{x}_i^{new}\theta - \mathbf{x}_i^{new}\hat{\theta})^2 \tag{21}$$

with respect to a LASSO constraint on θ. Formulae defining s and w can be found in Tran et al. (2012). 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.

References

Biau, G, Devroye, L. and Lugosi, G. (2008). Consistency of Random Forests and Other Averaging Classifiers. *Journal of Machine Learning Research*, **9**, 2015–2033.

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J., Sturmer, T. (2006). Variable selection for propensity score models. Am. J. Epidemiol. 163, 1149 – 1156.

Brookhart, M.A. and van der Laan, M.J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Computational Statistics and Data Analysis* **50**, 475 – 498.

Bühlmann, P., Kalisch, M. and Maathuis, M.H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* **97**, 261-278.

Efron, B., Hastie, T., Johnstone, T., and Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics* **32**, 407–499.

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. JAMA. 1996 Sep 18;276(11):889-97.

Crainiceanu, C., Dominici, F. and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* **95**, 635 – 651.

Fan, J., and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Freidman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* **41**, 361 – 372.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**, 320 – 328.

George, E. and McCulloch, R. E. (1993). Variable selection with Gibbs sampling. *Journal of the American Statistical Association* **88**, 881 – 889.

Hirano, K. and Imbens, G. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services and Outcomes Research Methodology. Volume 2, Numbers 3-4 (2001), 259-278, DOI: 10.1023/A:1020371312283.

Holland, P. 1986. Statistics and causal inference (with discussion). Journal of the American Statistical Association 81, 945 – 970.

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. Ann Intern Med. 1995 Feb 1;122(3):191-203.

Little, R. J. A. and Rubin, D. B. (2002). *The Statistical Analysis of Missing Data, 2nd edition.* New York: Wiley and Sons.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23**, 2937 – 2960.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference.* Cambridge University Press.

Chen Q, Wang S. Variable selection for multiply-imputed data with application to dioxin exposure study. Stat Med. 2013 Mar 25. doi: 10.1002/sim.5783. [Epub ahead of print] PubMed PMID: 23526243.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J. V. and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology* **27**, 85 – 95.

Robins, J. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics, Theory and Methods* **23**, 23792412.

Robins, J. M., Hernán, M. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550 – 560.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York : John Wiley.

Sato, T. and Matsuyama, Y. (2003). Marginal structural models as a tool for standardization. *Epidemiology* **14**, 680 – 686.

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

Tran, M. N., Nott, D. J. and Leng, C. (2012). The predictive Lasso. *Statistics and Computing* **22**, 1069 – 1084.

Van Buuren, S.(2012). *Flexible Imputation of Missing Data.* Boca Raton, FL: Chapman & Hall/CRC Press.

Van der Laan, M. J. and Rose, K. (2011). *Target Learning: Causal Inference for Observational and Experimental Data.* New York: Springer.

Vansteelandt, S., Bekaert, M. and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* **21**, 7-30.

Wang, C., Parmigiani, G., and Dominici, F. (2012) Bayesian effect estimation accounting for adjustment uncertainty (with discussion). *Biometrics* **68**, 661 – 676.

Wood, A. M., White, I. R. and Royston, P. (2008). How should variable selection be performed with mutiply imputed data? *Statistics in Medicine* **27**, 3227 – 3246.

Yang, X., Belin, T. R. and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61**, 498-506.