Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study.

Leo Guelman[a,*], Montserrat Guillén[b], Ana M. Pérez-Marín[b]

[a] *Royal Bank of Canada, RBC Insurance, 6880 Financial Drive, Mississauga, Ontario L5N 7Y5, Canada*
[b] *Dept. Econometrics, Riskcenter, University of Barcelona, Diagonal 690, E-08034 Barcelona, Spain*

Abstract

In many important settings, subjects can show significant heterogeneity in response to a stimulus or "treatment". For instance, a treatment that works for the overall population might be highly ineffective, or even harmful, for a subgroup of subjects with specific characteristics. Similarly, a new treatment may not be better than an existing treatment in the overall population, but there is likely a subgroup of subjects who would benefit from it. The notion that "one size may not fit all" is becoming increasingly recognized in a wide variety of fields, ranging from economics to medicine. This has drawn significant attention to personalize the choice of treatment, so it is optimal for each individual. An optimal personalized treatment is the one that maximizes the probability of a desirable outcome. We call the task of learning the optimal personalized treatment *personalized treatment learning*. From the statistical learning perspective, this problem imposes some challenges, primarily because the optimal treatment is unknown on a given training set. A number of statistical methods have been proposed recently to tackle this problem. However, to the best of our knowledge, there has been no attempt so far to provide a comprehensive view of these methods and to benchmark their performance. The purpose of this paper is twofold: i) to describe seven recently proposed methods for personalized treatment learning and compare their performance on an extensive numerical study, and ii) to propose a novel method labeled *causal conditional inference trees* and its natural extension to *causal conditional inference forests*. The results show that our new proposed method often outperforms the alternatives on the numerical settings described in this article. We also illustrate an application of the proposed method using data from a large Canadian insurer for the purpose of selecting the best targets for cross-selling an insurance product.

Keywords: personalized treatment learning, causal inference, marketing interventions

1. Introduction

In the past two decades, the rapid advances in data collection and storage technology have created vast quantities of data. The field of statistics was revolutionized from the development of algorithmic and data models (Brieman, 2001) in response to new challenging problems coming from science and industry, mostly resulting from the increasing size and complexity in the data structures. In this context, the concept of *learning from data* (Abu-Mostafa et al., 2012) has emerged as the task of extracting "implicit, previously unknown, and potentially useful information from data" (Frawley, 1991). A usual distinction is made between *supervised* and *unsupervised* learning. In the former, the objective is to predict the value of a response variable based on a collection of *observable* covariates. In the later, there is no response variable to "supervise" the learning process, and the objective is to find structures and patterns among the covariates.

In many important settings, the values of some covariates are not only observable, but they can be chosen at the discretion of a decision maker (Žliobaitė and Pechenizkiy, 2010). For instance, a doctor can choose the medical treatment for a patient among a set of alternatives, a company can decide the type of marketing intervention activity (direct mail, phone call, email, etc.) to make an offer to a client, a bank can decide the credit limit amount to offer to a client on a credit card. In all these examples, the objective is not necessarily to predict a response variable with high accuracy, but to select the optimal action or "treatment" for each subject based on his or her individual characteristics[1]. Optimal is understood here as the treatment that maximizes the probability of a desirable outcome. We call the task of learning the optimal personalized treatment *personalized treatment learning*.

A key challenge with personalized treatment learning is that the quantity we are trying to predict (i.e., the optimal personalized treatment) is unknown on a given training data set. As

*Corresponding author. Tel. : +1 905 606 1175; Fax: +1 905 286 4756
Email addresses: leo.guelman@rbc.com (Leo Guelman), mguillen@ub.edu (Montserrat Guillén), amperez@ub.edu (Ana M. Pérez-Marín)

[1]Domain knowledge can play an important role in the preliminary determination of the relevant subject characteristics to select the optimal treatment (Sinha and Zhao, 2008).

each subject can only be exposed to a single treatment, the value of the response under alternative treatments is unobserved; a problem also known as *the fundamental problem of causal inference* (Holland, 1986). This aspect makes this problem unique within the discipline of learning from data.

The underlying motivation for personalized treatment learning is that subjects can show significant heterogeneity in response to treatments, so making an accurate treatment choice for each subject becomes essential. For instance, a new treatment may not be better than an existing treatment in the overall population, but it might be beneficial/harmful for a subgroup of subjects. The idea that "one size may not fit all" has been increasingly recognized in a variety of disciplines, ranging from economics to medicine. Alemi et al. (2009) argue that improved statistical methods are needed for personalized treatments and proposed an adapted version of the K-Nearest-Neighbor (KNN) classifier (Cover and Hart, 1967). Imai and Ratkovic (2012) proposed a method that adapts the *support vector machine* classifier (Vapnik, 1995) and then apply it to a widely known dataset pertaining to the National Supported Work program (LaLonde, 1986; Dehejia and Wahba, 1999) to identify the characteristics of workers who greatly benefit from (or are negatively affected by) a job training program. Tian et al. (2012) proposed a method designed to deal with high dimensional covariates and they use it to identify breast cancer patients who may or may not benefit from a specific treatment based on the individual patient's gene expression profile. Liang et al. (2006) describe a web-based intervention support system to provide tailored interventions to individual patients with chronic diseases. Xu et al. (2008) proposed a Bayesian network model that integrates with other components to better support personalized mobile advertising applications. In the context of insurance, Guelman et al. (2012, 2013) proposed a method based on an adapted version of *random forests* to identify policyholders who are positively/negatively impacted from a client retention program. Also, Guelman and Guillén (2014) describe a framework to determine the optimal rate change (i.e., playing the role of the treatment) for each individual policyholder for the purpose of maximizing the overall expected profitability of an insurance portfolio.

In addition to the methods discussed above, other methods have been proposed in the literature, mostly in the context of clinical trials and direct marketing (Su et al., 2009; Qian and Murphy,

2011; Zhao et al., 2012; Jaśkowski and Jaroszewicz, 2012; Larsen, 2009; Radcliffe and Surry, 2011; Rubin and Waterman, 2006; Tang et al., 2013). However, to the best of our knowledge, there has been no attempt so far to provide a comprehensive view of the methods and to benchmark their performance. The purpose of this paper is twofold. First, to describe seven recently proposed methods for personalized treatment learning and compare their performance on synthetic data. While other methods beyond those covered in this article have been proposed, we believe the ones described here provide a good representation of the current state of methods designed for personalized treatment learning. Second, to propose a novel method labeled *causal conditional inference trees* and its natural extension to *causal conditional inference forests*.

This paper is organized as follows. Section 2 defines the scope of the personalized treatment learning problem. Section 3 follows with a detailed description of seven existing methods to tackle this problem. In Section 4, we discuss our new proposed method. In Section 5, we provide the finite sample performance of all methods under an extensive simulation study. The results show that our new proposed method often outperforms the alternatives on the numerical settings described in this article. Finally, in Section 6 we describe an empirical application of the proposed method, using data from a major Canadian insurer, to determine which Auto insurance policyholders are more likely to be positively stimulated to buy a Home policy as a result of a marketing cross-sell intervention activity.

2. Problem formulation

We postulate the *personalized treatment learning* problem in the context of Rubin's model of causality (Rubin, 1974, 1977, 1978, 2005). Under this model, we conceptualize the learning problem in terms of the potential outcomes under treatment alternatives, only one of which is observed for each subject. The causal effect of a treatment on a subject is defined in terms of the difference between an observed outcome and its counterfactual. The notation introduced below will be used throughout the paper.

In the following, we use upper case letters to denote random variables and lower case letters to denote values of the random variables. Assume that a sample of subjects is randomly assigned

to two treatment arms, denoted by A, $A \in \{0,1\}$, also referred as control and treatment states, respectively. Let $Y(a) \in \{0,1\}$ denote a binary potential outcome of a subject if assigned to treatment $A = a$, $a = \{0,1\}$. The observed outcome is $Y = AY(1) + (1-A)Y(0)$. Throughout this paper we assume a value of $Y = 1$ is more desirable than $Y = 0$. Each subject is characterized by a p-dimensional vector of baseline covariates $\mathbf{X} = (X_1, \ldots, X_p)^\top$. We assume the data consists of L independent and identically distributed realizations of $(Y, A, \mathbf{X})$, $\{(Y_\ell, A_\ell, \mathbf{X}_\ell),\ \ell = 1, \ldots, L\}$.

Under the assumption of randomization, treatment assignment A ignores its possible impact on the outcomes $Y(0)$ and $Y(1)$, and hence they are independent – i.e., using the notation of Dawid (1979), $\{Y_\ell(0), Y_\ell(1) \perp A_\ell\}$. In this context, the *average treatment effect* (ATE) can be estimated by

$$\begin{aligned} \tau &= E[Y_\ell(1) - Y_\ell(0)] \\ &= E[Y_\ell | A_\ell = 1] - E[Y_\ell | A_\ell = 0]. \end{aligned} \tag{1}$$

In observational studies, subjects assigned to different treatment conditions are not exchangeable and thus direct comparisons can be misleading (Rosenbaum and Rubin, 1983).

In many circumstances, subjects can show significant heterogeneity in response to treatments, in which case the ATE is of limited value. The problem addressed in this paper is the identification of subgroups of subjects for which the treatment is most beneficial (or most harmful) within the context of experimental data. As discussed by Holland and Rubin (1988), the most granular level of causal inference is the *individual treatment effect* (ITE), defined by $Y_\ell(1) - Y_\ell(0)$ for each subject $\ell = \{1, \ldots, L\}$. However, this is an unobserved quantity, as a subject is never observed simultaneously in both treatment states. The best approximation to the ITE that is possible to obtain in practice is the *subpopulation treatment effect* (STE), which is defined for a subject with individual covariate profile $\mathbf{X}_\ell = \mathbf{x}$ by

$$\begin{aligned}\tau(\mathbf{x}) &= E[Y_\ell(1) - Y_\ell(0)|\mathbf{X}_\ell = \mathbf{x}] \\ &= E[Y_\ell|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 1] - E[Y_\ell|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 0]. \end{aligned} \tag{2}$$

Understanding the precise nature of the STE variability can be extremely valuable to personalize the choice of treatment, so that it is most appropriate for each individual. Henceforth in this paper, we use the term *personalized treatment effect* (PTE) to refer to the subpopulation treatment effect (2).

A *personalized treatment rule* $\mathcal{H}$ is a map from the space of baseline covariates $\mathbf{X}$ to the space of treatments A, $\mathcal{H}(\mathbf{X}) : R^p \to \{0, 1\}$. An *optimal treatment rule* is the one that maximizes the expected outcome if the personalized treatment rule is implemented for the whole population[2], i.e., $E[Y(\mathcal{H}(\mathbf{X}))]$. A straightforward calculation gives the optimal personalized treatment rule $\mathcal{H}^* = \text{argmax}_{\mathcal{H}}\, E[Y(\mathcal{H}(\mathbf{X}))]$ for a subject with covariates $\mathbf{X}_\ell = \mathbf{x}$ as $\mathcal{H}^* = 1$ if $\tau(\mathbf{x}) > 0$, and $\mathcal{H}^* = 0$ otherwise. In many situations, the alternative treatments have unequal costs, in which case the decision rule can simply be replaced by $\mathcal{H}^* = 1$ if $\tau(\mathbf{x}) > c$, and $\mathcal{H}^* = 0$ otherwise, for some constant probability threshold c.

3. Models for personalized treatment learning

In this section, we describe seven of the most prominent methods discussed in the literature to estimate personalized treatment effects (PTE).

3.1. Indirect estimation methods

We discuss below three indirect methods for estimating personalized treatment effects. We labeled these methods "indirect" as they propose a systematic 2-stage procedure to estimate the PTE. In the first stage, they attempt to achieve high accuracy in predicting the outcome Y condi-

[2]Notice that since Y is binary, this expectation has a probabilistic interpretation. That is, $E[Y(\mathcal{H}(\mathbf{X}))] = \text{Prob}\Big(Y(\mathcal{H}(\mathbf{X})) = 1\Big)$.

tional on the covariates $\mathbf{X}$ and treatment A. In the second stage, they subtract the predicted value of Y under each treatment to obtain a PTE estimate.

The first method is the *probability decomposition model* discussed by Larsen (2009). The most intuitive approach to estimate personalized treatment effects is to fit two independent models for the response Y, one based on the treated subjects, $E[Y|\mathbf{X}, A = 1]$, and one based on the control subjects, $E[Y|\mathbf{X}, A = 0]$. An estimate of the PTE for a subject with covariate $\mathbf{X}_\ell = \mathbf{x}$ is then obtained by subtracting the estimated value of the response from the two models. Any conventional statistical or algorithmic binary classification method may serve to fit the models.

Alternatively, a second method in the same spirit is the *interaction* approach proposed by Lo (2002). This method consists in fitting a single model to the response on the main effects and adding interaction terms between each covariate $\mathbf{X} = (X_1, \ldots, X_p)^\top$ and the treatment indicator A. If the model is fitted using standard logistic regression, the estimated parameters of the interaction terms measure the additional effect of each covariate due to treatment. An estimate of the PTE for a subject with covariates $\mathbf{X}_\ell = \mathbf{x}$ is obtained by subtracting the predicted probabilities by setting, in turn, $A_\ell = 1$ and $A_\ell = 0$ in the fitted model.

The interaction method represents an improvement over the probability decomposition model approach in that it provides a formal means of performing significance tests of the interaction parameters between the treatment and the covariates. However, it suffers from overfitting problems when including all interaction effects with a high dimensional covariate space Zhao and Zeng (2012). Although, overfitting may be prevented by using LASSO logistic regression (Tibshirani, 1996) for variable selection and shrinkage, this method places the same LASSO constraints over main and treatment heterogeneity parameters. This may be problematic as the variability in the response attributable to the interaction effects is usually a small fraction of the variability in the response attributable to the main effects. To address this problem, a third method proposed by Imai and Ratkovic (2012), called *L2-SVM*, is an adapted version of the support vector machine (SVM) classifier (Vapnik, 1995). The SVM can be expressed as a penalization method (Hastie et al., 2009, p. 426) and this can be adapted to include separate LASSO constraints over the main and treatment heterogeneity parameters. Specifically, let $Y_\ell^* = 2Y_\ell - 1 \in \{-1, 1\}$ and consider the

following optimization problem

$$\min_{(\alpha,\theta)} \sum_{\ell=1}^{L} \left|1 - Y_\ell^*(\mu + \alpha^\top \mathbf{X}_\ell + \theta^\top \mathbf{X}_\ell A_\ell)\right|_+^2 + \lambda_{\mathbf{X}} \sum_{j=1}^{p} |\alpha_j| + \lambda_{\mathbf{XA}} \sum_{j=1}^{p} |\theta_j| \tag{3}$$

where $\lambda_{\mathbf{X}}$ and $\lambda_{\mathbf{XA}}$ are pre-specified separate LASSO penalties for the main effect parameters α and treatment heterogeneity parameters θ, respectively, $|x|_+ \equiv \max(x, 0)$ is the hinge-loss (Wahba, 2002), and μ is a constant term.

After model (3) is estimated, a PTE estimate can be obtained as follows. Let $\hat{R}_\ell = \hat{\mu} + \hat{\alpha}^\top \mathbf{X}_\ell + \hat{\theta}^\top \mathbf{X}_\ell A_\ell$ and $\hat{R}_\ell^*$ denote the predicted value $\hat{R}_\ell$ truncated at positive and negative one. The PTE is estimated as the difference in the truncated values of the predicted response under each treatment condition. That is,

$$\hat{\tau}(\mathbf{x}) = 1/2\Big[(\hat{R}_\ell | \mathbf{X} = \mathbf{x}_\ell, A_\ell = 1) - (\hat{R}_\ell | \mathbf{X} = \mathbf{x}_\ell, A_\ell = 0)\Big]. \tag{4}$$

A key problem with the indirect estimation methods is the mismatch between the target variable they attempt to estimate and the target variable defined in (2). For instance, even when any of the indirect estimation methods are correctly specified to predict Y_ℓ conditional on covariates $\mathbf{X}_\ell = \mathbf{x}$ and treatment A, it is not guaranteed that these models can accurately predict $Y_\ell(1) - Y_\ell(0)$ conditional on the same covariates. This is because these methods emphasize the prediction accuracy on the response, not the accuracy in estimating the change in the response *caused* by the treatment at the subject level.

3.2. Modified covariate method

This method, proposed by Tian et al. (2012), consists in modifying the covariates in a simple way, and then fitting an appropriate regression model using the modified covariates. A key advantage of this approach is that it avoids having to directly model the main effects.

Specifically, this method involves performing the following steps: i) transform the treatment indicator as $A_\ell^* = 2A_\ell - 1 \in \{-1, 1\}$, ii) transform each covariate in $\mathbf{X}_\ell$ as $\mathbf{Z}_\ell = \mathbf{X}_\ell^* A_\ell^*/2$, where $\mathbf{X}^*$ is the centered version of $\mathbf{X}$, and iii) fit a regression model to predict the outcome variable Y on the modified covariates $\mathbf{Z}$. For instance, using a logistic regression model, estimate

$$P(Y = 1|\mathbf{X}, A) = \frac{\exp(\gamma^\top \mathbf{Z})}{1 + \exp(\gamma^\top \mathbf{Z})}. \tag{5}$$

Under the very general assumption that $P(A^* = 1) = P(A^* = -1) = 1/2$, a surrogate to the personalized treatment effect for a subject with covariates $\mathbf{X}_\ell = \mathbf{x}$ is given by

$$\hat{\tau}(\mathbf{x}) = \frac{\exp(\hat{\gamma}^\top \mathbf{x}/2) - 1}{\exp(\hat{\gamma}^\top \mathbf{x}/2) + 1}. \tag{6}$$

To see that (6) is an appropriate estimate, we must consider the maximum likelihood estimator of model (5). It is easy to see (Tian et al., 2012) that the minimizer of $E\{Yf(\mathbf{X})A^* - \log(1 + \mathbf{exp}(f(\mathbf{X})A^*))\}$, with $f(\mathbf{X}) = \gamma^\top \mathbf{X}^*/2$, is given by

$$f^*(\mathbf{x}) = \log\Big\{\frac{1 + \tau(\mathbf{x})}{1 - \tau(\mathbf{x})}\Big\}, \tag{7}$$

where $\tau(\mathbf{x})$ is the personalized treatment effect defined in (2). Therefore, (6) may serve as an estimate of the PTE.

In case the dimension of $\mathbf{X}$, p, is high, appropriate variable selection procedures can directly be applied to the modified data. For instance, an *L1*-regularized logistic regression (Hastie et al., 2009, p. 125) can be estimated by minimizing $\frac{1}{L}\sum_{\ell=1}^{L} -\{Y_\ell \gamma^\top \mathbf{Z}_\ell - \log(1 + \exp(\gamma^\top \mathbf{Z}_\ell)) + \lambda_0 \sum_{j=1}^{p} |\gamma_j|\}$, where λ_0 is a pre-specified LASSO penalty.

In the derivation above, the assumption of equal probability of treatments is used. This may be perceived as very restrictive as this assumption is unlikely to hold in practice. However, various resampling methods from the machine learning literature (Weiss and Provost, 2003; Estabrooks et al., 2004; Chawla, 2005) could be used for the purpose of balancing treatments.

3.3. Modified outcome method

The modified outcome method was proposed by Jaśkowski and Jaroszewicz (2012). It consists in first defining a new outcome variable W such that

$$W_\ell = \begin{cases} 1 & \text{if } A_\ell = 1 \text{ and } Y_\ell = 1 \\ 1 & \text{if } A_\ell = 0 \text{ and } Y_\ell = 0 \\ 0 & \text{otherwise,} \end{cases}$$

and then fitting a binary regression model to W on the baseline covariates $\mathbf{X}$. Recall that we assumed that a value of $Y = 1$ is more desirable that $Y = 0$, thus we can intuitively think of $W = 1$ as the event of obtaining a potential outcome under treatment which is at least as good as the observed outcome. The probability of this event is given by

$$\begin{aligned} P(W_\ell = 1|\mathbf{X}_\ell = \mathbf{x}) &= P(W_\ell = 1|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 1)P(A_\ell = 1|\mathbf{X}_\ell = \mathbf{x}) + \\ &\quad P(W_\ell = 1|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 0)P(A_\ell = 0|\mathbf{X}_\ell = \mathbf{x}) \\ &= P(Y_\ell = 1|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 1)P(A_\ell = 1|\mathbf{X}_\ell = \mathbf{x}) + \\ &\quad P(Y_\ell = 0|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 0)P(A_\ell = 0|\mathbf{X}_\ell = \mathbf{x}) \\ &= P(Y_\ell = 1|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 1)P(A_\ell = 1) + \\ &\quad P(Y_\ell = 0|\mathbf{X}_\ell = \mathbf{x}, A_\ell = 0)P(A_\ell = 0), \end{aligned}$$

where the last equality follows from the randomization assumption. Now, making the same assumption as in the modified covariate method that $P(A = 1) = P(A = 0) = 1/2$ we obtain

$$\begin{aligned} \tau(\mathbf{x}) &= P(Y_\ell = 1|A_\ell = 1, \mathbf{X}_\ell = \mathbf{x}) - P(Y_\ell = 1|A_\ell = 0, \mathbf{X}_\ell = \mathbf{x}) \\ &= 2P(W_\ell = 1|\mathbf{X}_\ell = \mathbf{x}) - 1. \end{aligned}$$

Hence, if for instance a logistic regression model is used to estimate

$$P(W = 1|\mathbf{X}, A) = \frac{\exp(\beta^\top \mathbf{X})}{1 + \exp(\beta^\top \mathbf{X})}, \tag{8}$$

then

$$\hat{\tau}(\mathbf{x}) = 2\frac{\exp(\hat{\beta}^\top \mathbf{X})}{1 + \exp(\hat{\beta}^\top \mathbf{X})} - 1 \tag{9}$$

can be used as a surrogate to the PTE.

In the Appendix, we show that the maximum likelihood estimator (MLE) of the working models (8) and (5) are equivalent and so they produce similar PTE estimates.

3.4. Causal K-nearest neighbor (CKNN)

A simple non-parametric method briefly discussed by Alemi et al. (2009) and also by Su et al. (2012) to estimate personalized treatment effects is to use a modified version of the K-Nearest-Neighbor (KNN) classifier (Cover and Hart, 1967).

The basic idea of the CKNN algorithm is that to estimate the personalized treatment effect for a target subject, we may wish to weight the evidence of subjects similar to the target more heavily. Consider a subject with covariates $\mathbf{X}_\ell = \mathbf{x}$ and a neighborhood of $\mathbf{x}$, $\mathcal{S}_k(\mathbf{x})$, represented by a sphere centered at $\mathbf{x}$ containing precisely K subjects, independently of their outcome Y and treatment type A. An estimate of the PTE is given by

$$\hat{\tau}(\mathbf{x}) = \frac{\sum_{\ell:\mathbf{x}_\ell \in \mathcal{S}_k(\mathbf{x})} Y_\ell A_\ell}{\sum_{\ell:\mathbf{x}_\ell \in \mathcal{S}_k(\mathbf{x})} A_\ell} - \frac{\sum_{\ell:\mathbf{x}_\ell \in \mathcal{S}_k(\mathbf{x})} Y_\ell (1 - A_\ell)}{\sum_{\ell:\mathbf{x}_\ell \in \mathcal{S}_k(\mathbf{x})} (1 - A_\ell)}. \tag{10}$$

The CKNN approach proposed in (10) assigns an equal weight of 1 to each of the K subjects within the neighbor $\mathcal{S}_k(\mathbf{x})$ and 0 weight to all other subjects. Alternatively, it is common to use kernel smoothing methods to assign weights that die off smoothly with the distance $||\mathbf{x}_\ell - \mathbf{x}||$ for all subjects $\ell = \{1, \ldots, L\}$. Also, notice that (10) is defined if at least one control and one treated subject are in the neighbor of $\mathbf{x}$ ($K \geq 2$).

A severe limitation of this method is that the entire training data have to be stored to score new subjects, leading to expensive computations for large data sets.

3.5. Uplift random forests

Tree-based models represent an intuitive approach to estimate (2), as appropriate split criteria can be designed to partition the input space into subgroups with heterogeneous treatment effects.

Uplift random forests is a tree-based method proposed by Guelman et al. (2013) to estimate personalized treatment effects. Algorithm 1 shows the details. In short, an ensemble of B trees are grown, each built on a fraction ν of the training data[3] (which includes both treatment and control records). The sampling, motivated by Friedman (2002), incorporates randomness as an integral part of the fitting procedure. This not only reduces the correlation between the trees in the sequence, but also reduces the computing time by the same fraction ν. A typical value for ν can be 1/2, although for large data, it can be substantially smaller. The tree-growing process involves selecting $n \leq p$ covariates at random as candidates for splitting. This adds an additional layer of randomness, which further reduces the correlation between trees, and hence reduces the variance of the ensemble. The split rule is based on a measure of distributional divergence, as defined in Rzepakowski and Jaroszewicz (2012), also discussed below. The individual trees are grown to maximal depth (i.e., no pruning is done). The estimated personalized treatment effect is obtained by averaging the predictions of the individual trees in the ensemble.

As the fundamental idea is to maximize the distance in the class distributions of the response Y between treatment and control groups, it is sensible to construct a split criteria by borrowing the concept of distributional divergence from information theory. In particular, if we let $P\{Y(1)\}$ and $P\{Y(0)\}$ be the class probability distributions over the response variable Y for the treatment and control, respectively, then *Kullback–Leibler distance* (KL) or *Relative Entropy* (Cover and Thomas, 1991, p. 9) between the two distributions is given by

$$KL\Big(P\{Y(1)\}||P\{Y(0)\}\Big) = \sum_{Y(A)\in\{0,1\}} P\{Y(1)\} \log\frac{P\{Y(1)\}}{P\{Y(0)\}}, \tag{11}$$

where the logarithm is to base two. The Kullback–Leibler distance is always nonnegative and

[3]In the standard random forest algorithm, bootstrap samples of the training data are drawn before fitting each tree. Our motivation for sampling a fraction of the data instead, was to reduce computational time on large data sets.

is zero if and only if $P\{Y(1)\} = P\{Y(0)\}$. Since the KL distance is non-symmetric, it is not a true distance measure. However, it is frequently useful to think of KL as a measure of divergence between distributions.

For any node, suppose there is a candidate split Ω which divides it into two child nodes, n_L and n_R, denoting the left and right node respectively. Further let L be the total number of subjects in the parent node and suppose L_{n_L} and L_{n_R} represent the number of subjects that go into n_L and n_R, respectively. Conditional on a split Ω, distributional divergence can be expressed as the KL distance, weighted by the proportion of subjects in each node

$$KL\Big(P\{Y(1)\}||P\{Y(0)\}|\Omega\Big) = \frac{1}{L}\sum_{i\in\{n_L,n_R\}} L_i KL\Big(P\{Y(1)|i\}||P\{Y(0)|i\}\Big). \tag{12}$$

Now, define KL_{gain} as the increase in the KL distance from a split Ω, relative to the KL distance in the parent node

$$KL_{gain}(\Omega) = KL\Big(P\{Y(1)\}||P\{Y(0)\}|\Omega\Big) - KL\Big(P\{Y(1)\}||P\{Y(0)\}\Big). \tag{13}$$

The final splitting rule adds a normalization factor to (13). This factor attempts to penalize splits with unbalanced proportions of subjects associated with each child node, as well as splits that result in unbalanced treatment/control proportion in each child node (since such splits are not independent of the group assignment). The final split criterion is then given by

$$KL_{ratio}(\Omega) = \frac{KL_{gain}(\Omega)}{KL_{norm}(\Omega)} \tag{14}$$

where

$$KL_{norm}(\Omega) = H\Big(\frac{L(1)}{L}, \frac{L(0)}{L}\Big)KL\Big(P\{\Omega(1)\}||P\{\Omega(0)\}\Big) + \frac{L(1)}{L}H\Big(P\{\Omega(1)\}\Big) + \frac{L(0)}{L}H\Big(P\{\Omega(0)\}\Big). \tag{15}$$

$L(A)$ in (15) denotes the number of subjects in treatment $A \in \{0, 1\}$, $P\{\Omega(A)\}$ represents the

probability distribution over the split outcomes $\{n_L, n_R\}$ for subjects with treatment A, and $H(.)$ is the *entropy* function, defined by $H(P\{\Omega(A)\}) = -P\{\Omega(A) = n_L\}\log(P\{\Omega(A) = n_L\}) - P\{\Omega(A) = n_R\}\log(P\{\Omega(A) = n_R\})$ and $H(\frac{L(1)}{L}, \frac{L(0)}{L}) = -\frac{L(1)}{L}\log(\frac{L(1)}{L}) - \frac{L(0)}{L}\log(\frac{L(0)}{L})$.

The last two terms in (15) penalize splits with a large number of outcomes, by means of the sum of entropies of the split outcomes in treatment and control groups weighted by the proportion of training cases in each group. The first term penalizes uneven splits, which is measured by the divergence in the distribution of the split outcomes between the groups. This term is multiplied by the entropy of the proportion of instances in treatment and control groups. This is to explicitly impose a smaller penalty when there is not enough data in one of these groups.

A problem with the KL_{ratio} is that extremely low values of the KL_{norm} may favor splits despite their low KL_{gain}. To avoid this, the KL_{ratio} criterion selects splits that maximize the KL_{ratio}, subject to the constraint that the KL_{gain} must be at least as great as the average KL_{gain} over all splits considered.

Algorithm 1 Uplift random forest

for $b = 1$ to B **do**
  Sample a fraction ν of the training observations L without replacement
  Grow an uplift decision tree UT_b to the sampled data:
  for each terminal node **do**
    repeat
      Select n covariates at random from the p covariates
      Select the best variable/split-point among the n covariates based on KL_{ratio}
      Split the node into two branches
    until a minimum node size l_{min} is reached
  end for
end for
Output the ensemble of uplift trees UT_b; $b = \{1, \ldots, B\}$
The predicted personalized treatment effect for a new data point $\mathbf{x}$, is obtained by averaging the predictions of the individual trees in the ensemble: $\hat{\tau}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} UT_b(\mathbf{x})$

4. Causal conditional inference trees

We propose here a tree-based method to estimate personalized treatment effects, with important enhancements over the uplift random forest algorithm. There are two fundamental aspects in which uplift random forests could be significantly improved: overfitting and the selection bias towards

covariates with many possible splits. The development of the framework introduced here to tackle these issues was motivated by the *unbiased recursive partitioning* method proposed by Hothorn et al. (2006).

With regards to overfitting, recall that the individual trees in the forest are grown to maximal depth. While this helps to reduce bias, there is the familiar tradeoff with variance. In the context of personalized treatment effects, the overfitting problem is exacerbated as, generally, the variability in the response from the treatment heterogeneity effects is small relative to the variability in the response from the main effects. If the fitted model is not able to distinguish well between the relative strength of these two effects, that may easily translate into overfitting problems. In conventional decision trees (Brieman et al., 1984; Quinlan, 1993), overfitting is solved by a pruning procedure. This consists in traversing the tree bottom up and testing for each (non-terminal) node, whether collapsing the subtree rooted at that node with a single leaf would improve the model's generalization performance. Tree-based methods proposed in the literature to estimate personalized treatment effects (Rzepakowski and Jaroszewicz, 2012; Su et al., 2012; Radcliffe and Surry, 2011) use some sort of pruning. However, the pruning procedures used by these methods are all *ad hoc* and lack a theoretical foundation.

Besides the overfitting problem, the second concern is the biased variable selection towards covariates with many possible splits or missing values. This problem is also present in conventional decision trees, such as CART (Brieman et al., 1984) and C4.5 (Quinlan, 1993), and results from the maximization of the split criterion over all possible splits simultaneously (Kass, 1980; Brieman et al., 1984, p. 42).

Following the framework proposed by Hothorn et al. (2006), we improved considerably the generalization performance of the uplift random forest method by solving both the overfitting and the biased variable selection problems. The key to the solution is the separation between the variable selection and the splitting procedure, coupled with a statistically motivated and computational efficient stopping criteria based on the theory of permutation tests developed by Strasser and Weber (1999).

The pseudocode of the proposed algorithm is shown in Algorithm 2. The most relevant aspects

to discuss are steps 7-12. Specifically, for each terminal node in the tree we test the global null hypothesis of no interaction effect between the treatment A and *any* of the n covariates selected at random from the set of p covariates. The global hypothesis of no interaction is formulated in terms of n partial hypotheses $H_0^j : E[W|X_j] = E[W],\ j = \{1, \ldots, n\}$, with the global null hypothesis $H_0 = \cap_{j=1}^{n} H_0^j$, where W is defined as in the modified outcome method discussed in Section 3.3. Thus, a conditional independence test of W and X_j has a causal interpretation for the treatment effect for subjects with baseline covariate X_j. Multiplicity in testing can be handled via Bonferroni-adjusted P values or alternative adjustment procedures (Wright, 1992; Shaffer, 1995; Benjamini and Hochberg, 1995). When we are not able to reject H_0 at a prespecified significance level α, we stop the splitting process at that node. Otherwise, we select the j^*th covariate X_{j*} with the smallest adjusted P value. The algorithm then induces a partition Ω^* of the covariate X_{j*} in two disjoint sets $\mathcal{M} \subset X_{j*}$ and $X_{j*} \setminus \mathcal{M}$ based on the split criterion discussed below. This statistical approach prevents overfitting, without requiring any form of pruning or cross-validation.

One approach to measure the independence between W and X_j would be to use a classical statistical test, such as a Pearson's chi-squared. However, the assumed distribution from these tests is only a valid approximation to the actual distribution in the large-sample case, and this does not likely hold near the leaves of the decision tree. Instead, we measure independence based on the theoretical framework of permutation tests, which is admissible for arbitrary sample sizes. Strasser and Weber (1999) developed a comprehensive theory based on a general functional form of multivariate linear statistics appropriate for arbitrary independence problems. Specifically, to test the null hypothesis of independence between W and $X_j,\ j = \{1, \ldots, n\}$, we use linear statistics of the form

$$\mathcal{T}_j = \text{vec}\left(\sum_{\ell=1}^{L} g(X_{j\ell}) h(W_\ell, (W_1, \ldots, W_L))^\top \right) \in R^{u_j v \times 1} \tag{16}$$

where $g : X_j \to R^{u_j \times 1}$ is a transformation of the covariate X_j and $h : W \to R^{v \times 1}$ is called the *influence function*. The "vec" operator transforms the $u_j \times v$ matrix into a $u_j v \times 1$ column vector. The distribution of $\mathcal{T}_j$ under the null hypothesis can be obtained by fixing $X_{j1}, \ldots, X_{jL}$ and conditioning on all possible permutations S of the responses $W_1, \ldots, W_L$. A univariate test

statistic c is then obtained by standardizing $\mathcal{T}_j \in R^{u_j v \times 1}$ based on its conditional expectations $\mu_j \in R^{u_j v \times 1}$ and covariance $\Sigma_j \in R^{u_j v \times u_j v}$, as derived by Strasser and Weber (1999). A common choice is the maximum of the absolute values of the standardized linear statistic

$$c_{\max}(\mathcal{T}, \mu, \Sigma) = \max \frac{\mathcal{T} - \mu}{\text{diag}(\Sigma)^{1/2}}, \tag{17}$$

or a quadratic form

$$c_{\text{quad}}(\mathcal{T}, \mu, \Sigma) = (\mathcal{T} - \mu)\Sigma^{+}(\mathcal{T} - \mu)^{\top}, \tag{18}$$

where Σ^{+} is the Moore-Penrose inverse of Σ. Many well-known classical tests (e.g., Pearson's chi-squared, Cochran-Mantel-Haenszel, Wilcoxon-Mann-Whitney) can be formulated from (16) by choosing the appropriate transformation g, influence function h and test statistic c to map the linear statistic $\mathcal{T}$ into the real line. This sheds light on the extension of the proposed method to response variables measured in arbitrary scales and multi-category or continuos treatment settings.

In step 11 of Algorithm 2, we select the covariate X_{j*} with smallest adjusted P value. The P value P_j is given by the number of permutations $s \in S$ of the data with corresponding test statistic exceeding the observed test statistic $t \in R^{u_j v \times 1}$. That is,

$$P_j = P(c(\mathcal{T}_j, \mu_j, \Sigma_j) \geq c(t_j, \mu_j, \Sigma_j)|S).$$

For moderate to large samples sizes, it might not be possible to obtain the exact distribution (calculated exhaustively) of the test statistic. However, we can approximate the exact distribution by computing the test statistic from a random sample of the set of all permutations S. In addition, Strasser and Weber (1999) showed that the asymptotic distribution of the test statistic given by (17) tends to multivariate normal with parameters μ and Σ as $L \to \infty$. The test statistic (18) follows an asymptotic chi-squared distribution with degrees of freedom given by the rank of Σ. Therefore, asymptotic P values can be computed for these test statistics.

Once we select the covariate X_{j*} to split, we next use a split criteria which explicitly attempts

to find subgroups with heterogeneous treatment effects. Specifically, we use the following measure proposed by Su et al. (2009) and also implemented later by Radcliffe and Surry (2011) for assessing the personalized treatment effect from a split Ω

$$G^2(\Omega) = \frac{(L-4)\{(\bar{Y}_{n_L}(1) - \bar{Y}_{n_L}(0)) - (\bar{Y}_{n_R}(1) - \bar{Y}_{n_R}(0))\}^2}{\hat{\sigma}^2\{1/L_{n_L}(1) + 1/L_{n_L}(0) + 1/L_{n_R}(1) + 1/L_{n_R}(0)\}} \tag{19}$$

where n_L and n_R denotes the left and right child nodes, respectively, $L_{i\in\{n_L,n_R\}}(A)$ denotes the number of observations in child node i exposed to treatment $A \in \{0,1\}$, and

$$\bar{Y}_{i\in\{n_L,n_R\}}(1) = \frac{\sum_{\forall \ell \in i} Y_\ell A_\ell}{\sum_{\forall \ell \in i} A_\ell}, \tag{20}$$

$$\bar{Y}_{i\in\{n_L,n_R\}}(0) = \frac{\sum_{\forall \ell \in i} Y_\ell (1-A_\ell)}{\sum_{\forall \ell \in i} (1-A_\ell)}, \tag{21}$$

$$\hat{\sigma}^2 = \sum_{A\in\{0,1\}} \sum_{i\in\{n_L,n_R\}} L_i(A)\bar{Y}_i(A)(1-\bar{Y}_i(A)). \tag{22}$$

The best split is given by $G^2(\Omega^*) = \max_\Omega G^2(\Omega)$ – i.e., the split that maximizes the criterion $G^2(\Omega)$ among all permissible splits. It can easily be seen (Su et al., 2009), that the split criteria given in (19) is equivalent to a chi-squared test for testing the interaction effect between the treatment and the covariate X_{j*} dichotomized at the value given by the split Ω.

Algorithm 2 Causal conditional inference forests

for $b = 1$ to B **do**
  Draw a sample with replacement from the training observations L such that P(A=1) = P(A=0) = 1/2
  Grow a conditional causal inference tree $CCIT_b$ to the sampled data:
  for each terminal node **do**
    repeat
      Select n covariates at random from the p covariates
      Test the global null hypothesis of no interaction effect between the treatment A and any of the n covariates (i.e., $H_0 = \cap_{j=1}^{n} H_0^j$, where $H_0^j : E[W|X_j] = E[W]$) at a level of significance α based on a permutation test
      if the null hypothesis H_0 cannot be rejected **then**
        Stop
      else
        Select the j^*th covariate X_{j*} with the strongest interaction effect (i.e., the one with the smallest adjusted P value)
        Choose a partition Ω^* of the covariate X_{j*} in two disjoint sets $\mathcal{M} \subset X_{j*}$ and $X_{j*} \setminus \mathcal{M}$ based on the $G^2(\Omega)$ split criterion
      end if
    until a minimum node size l_{min} is reached
  end for
end for
Output the ensemble of causal conditional inference trees $CCIT_b;\ b = \{1, \ldots, B\}$
The predicted personalized treatment effect for a new data point $\mathbf{x}$, is obtained by averaging the predictions of the individual trees in the ensemble: $\hat{\tau}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} CCIT_b(\mathbf{x})$

5. Simulation studies

In this section, we conduct a numerical study for the purpose of assessing the finite sample performance of the analytical methods introduced in Sections 3 and 4. The *L2-SVM* (`l2svm`), Causal K-nearest neighbor (`cknn`), Uplift random forests (`upliftRF`) and Causal conditional inference forests (`ccif`) methods require specialized software. We developed software in the R statistical computing environment to implement the later three methods, and used the R package `FindIt` for the first one – developed by the authors of the method (Imai and Ratkovic, 2012). The remaining methods – i.e., the modified covariate method (`mcm`), modified outcome method (`mom`), probability decomposition model (`pdm`) and interaction method (`int`), can be implemented straightforwardly using readily available software.

Our simulation framework is based on the one described in Tian et al. (2012), but with a few

modifications. We evaluate the performance of the aforementioned methods in eight simulation settings, by varying i) the relative strength of the main effects relative to the treatment heterogeneity effects, ii) the degree of correlation among the covariates, and iii) the noise levels in the response.

We generated L independent binary samples from the regression model

$$Y = I\left(\Big[\sum_{j=1}^{p} \eta_j X_j + \sum_{j=1}^{p} \delta_j X_j A_j^* + \epsilon \Big] \geq 0 \right), \tag{23}$$

where the covariates $(X_1, \ldots, X_p)$ follow a mean zero multivariate normal distribution with covariance matrix $(1-\rho)\mathbf{I}_p + \rho \mathbf{1}^\top \mathbf{1}$, $A_\ell^* = 2A_\ell - 1 \in \{-1, 1\}$ was generated with equal probability at random, and $\epsilon \sim N(0, \sigma_0^2)$. We let $L = 200$, $p = 20$, and $(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \ldots, \delta_p) = (1/2, -1/2, 1/2, -1/2, 0, \ldots, 0)$.

Table 1 shows the simulation scenarios. The first four scenarios model a situation in which the variability in the response from the main effects is twice as big as that from the treatment heterogeneity effects, whereas in the last four scenarios the variability in the response from the main effects is four times as big as that from the treatment heterogeneity effects. Each of these scenarios were tested under none and moderate correlation among the covariates ($\rho = 0$ and $\rho = 0.5$), and two levels of noise ($\sigma_0 = \sqrt{2}$ and $\sigma_0 = 2\sqrt{2}$).

Table 1: Simulation scenarios

Scenario	η_j	ρ	σ_0
1	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0	$\sqrt{2}$
2	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0	$2\sqrt{2}$
3	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0.5	$\sqrt{2}$
4	$(-1)^{(j+1)}I(3 \leq j \leq 10)/2$	0.5	$2\sqrt{2}$
5	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0	$\sqrt{2}$
6	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0	$2\sqrt{2}$
7	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0.5	$\sqrt{2}$
8	$(-1)^{(j+1)}I(3 \leq j \leq 10)$	0.5	$2\sqrt{2}$

Note. This table displays the numerical settings considered in the simulations. Each scenario is parameterized by the strength of the main effects, η_j, the correlation among the covariates, ρ, and the magnitude of the noise, σ_0.

The key benefit of simulations in the context of personalized treatment effects is that the "true" treatment effect is known for each subject, a value which is not observed in empirical data. The performance of the analytical methods was measured using the *Spearman's rank correlation* coefficient between the estimated treatment effect $\hat{\tau}(X)$ derived from each model, and the "true" treatment effect

$$\begin{aligned}\tau(\mathbf{X}) &= E[Y(1) - Y(0)|\mathbf{X}] \\ &= P\left(\sum_{j=1}^{p}(\eta_j + \delta_j)X_j \leq \epsilon\right) - P\left(\sum_{j=1}^{p}(\eta_j - \delta_j)X_j \leq \epsilon\right) \\ &= F\left(\sum_{j=1}^{p}(\eta_j + \delta_j)X_j\right) - F\left(\sum_{j=1}^{p}(\eta_j - \delta_j)X_j\right), \end{aligned} \tag{24}$$

in an independently generated test set with a sample size of 10000. In (24), F denotes the cumulative distribution function of a normal random variable with mean zero and variance σ_0^2.

Variable selection for the `mcm`, `mom`, `pdm` and `int` methods was performed using the LASSO via a 10-fold cross-validation procedure. Based on this selection method, we found cases where the LASSO could not select any non-zero covariate based on cross-validation. Similarly to Tian et al.

(2012), in those cases we simply forced the correlation coefficient to be zero in the test set since the method did not find anything informative. For this reason, we alternatively fit these methods based on random forests (Breiman, 2001) using its default settings[4]. We refer to these methods based on random forest fits as `mcm-RF`, `mom-RF`, `pdm-RF` and `int-RF`. The optimal values for the LASSO penalties in (3) for the `l2svm` method, and the value of K in (10) for the `ccif` method, were also selected via 10-fold cross-validation. Lastly, the methods `upliftRF` and `ccif` were fitted using their default settings[5].

The results over a 100 repetitions of the simulation for the first and last four simulation scenarios are shown in Figures 1 and 2, respectively. These figures illustrate the boxplots of the Spearman's rank correlation coefficient between $\hat{\tau}(X)$ and $\tau(X)$. The boxplots within each simulation scenario are shown in increasing order of performance based on the average correlation. The `ccif` method performed either the best or next to the best in all eight scenarios.

6. An insurance cross-sell application

In this section, we apply the new proposed method to an insurance marketing application. The data used for this analysis is based on a direct mail campaign implemented by a large Canadian insurer between June 2012 and May 2013. The objective of the campaign was to drive more business from the existing portfolio of Auto Insurance clients by cross-selling them a Home Insurance policy with the company. The regular savings via the multi-product discount was prominently featured and positioned as the key element in the offer to the clients. In addition to the direct mail, clients were also contacted over the phone to further motivate them to initiate a Home policy quote. A randomized control group was also included as part of the campaign design, consisting of clients who

[4]Specifically, we fitted the models using $B = 500$ trees and $n = \sqrt{p}$ as the number of variables randomly sampled as candidates at each split.

[5]In both cases we used $B = 500$ trees and $n = p/3$ as the number of variables randomly sampled as candidates at each split. For `ccif` we set the P value $= 0.05$.

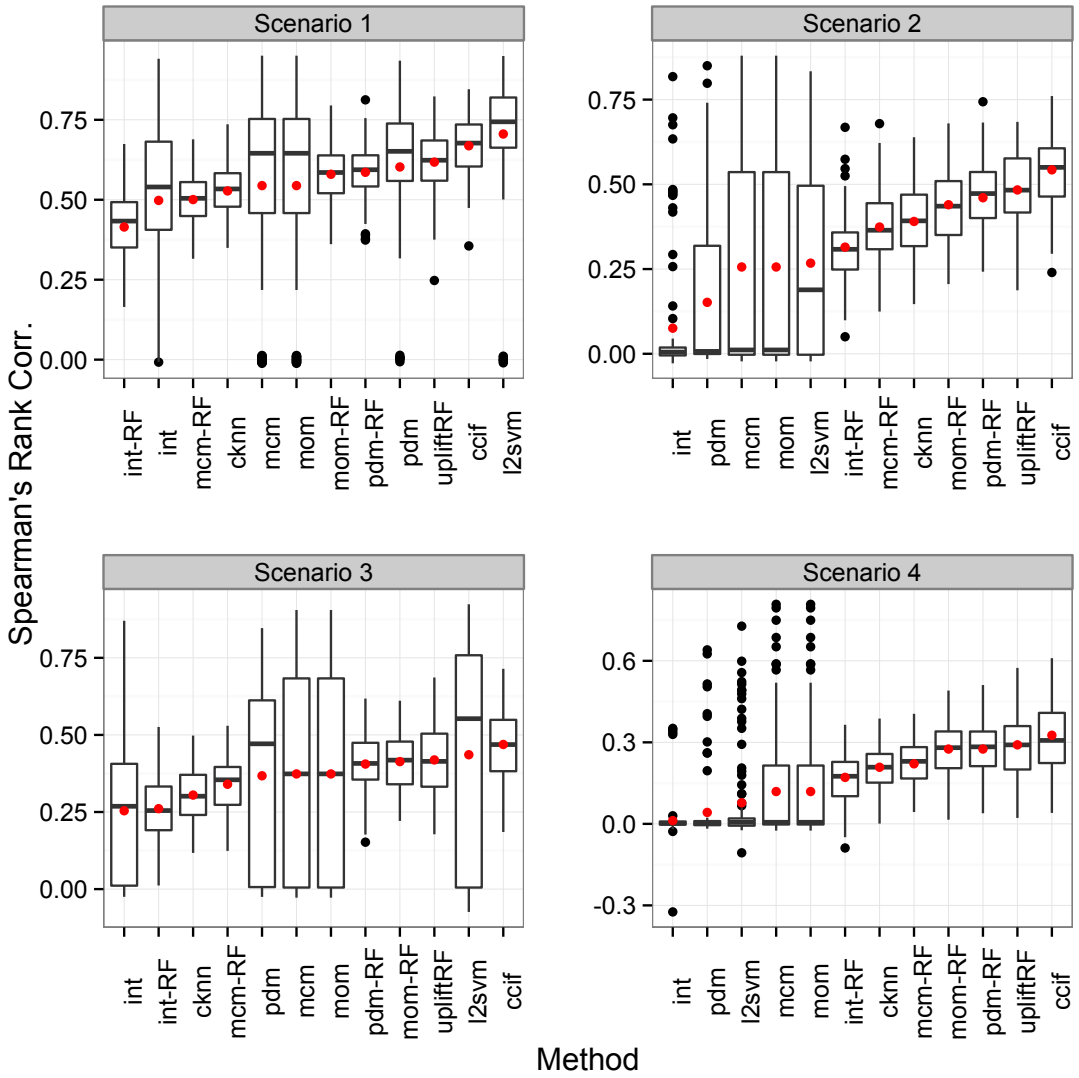

Figure 1: *Boxplots of the Spearman's rank correlation coefficient between the estimated treatment effect $\hat{\tau}(X)$ and the "true" treatment effect $\tau(X)$ for all methods. The plots illustrate the results for the **1-4** simulation scenarios, which model a situation with "**stronger**" treatment heterogeneity effects, under none and moderate correlation among the covariates ($\rho = 0$ and $\rho = 0.5$) and two levels of noise ($\sigma_0 = \sqrt{2}$ and $\sigma_0 = 2\sqrt{2}$). The boxplots within each simulation scenario are shown in increasing order of performance based on the average correlation.*

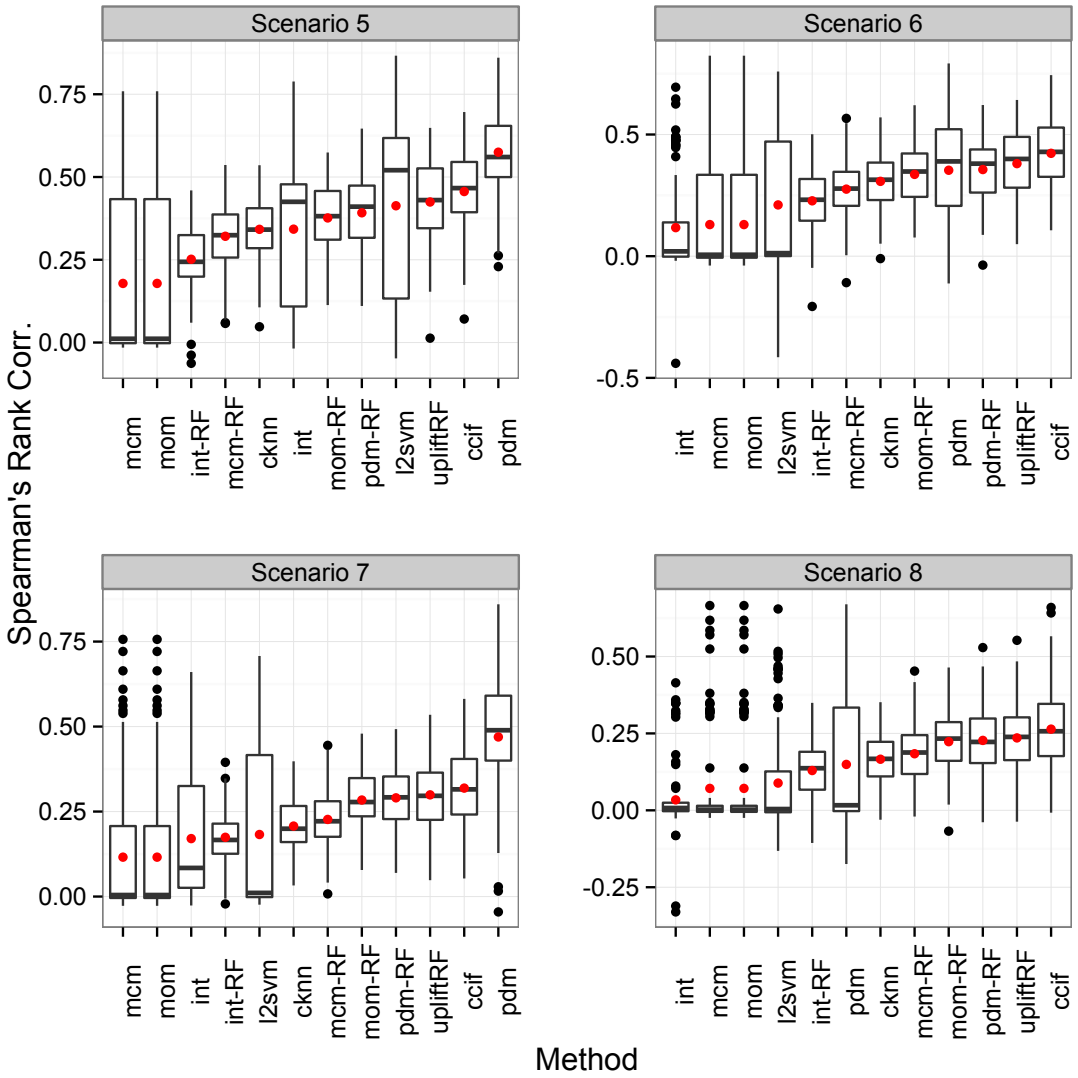

Figure 2: *Boxplots of the Spearman's rank correlation coefficient between the estimated treatment effect $\hat{\tau}(X)$ and the "true" treatment effect $\tau(X)$ for all methods. The plots illustrate the results for the* ***5-8*** *simulation scenarios, which model a situation with "**weaker**" treatment heterogeneity effects, under none and moderate correlation among the covariates ($\rho = 0$ and $\rho = 0.5$) and two levels of noise ($\sigma_0 = \sqrt{2}$ and $\sigma_0 = 2\sqrt{2}$). The boxplots within each simulation scenario are shown in increasing order of performance based on the average correlation.*

were not mailed or called. The response variable is determined by whether the client purchased the Home policy between the mail date and 3 months thereafter. In addition to the response, the dataset contains approximately 50 covariates related to the Auto policy, including driver and vehicle characteristics and general policy information.

Table 2 shows the cross-sell rates by group. The average treatment effect of **0.34%** (2.55% - 2.21%) is not statistically significant with a P value of 0.23 based on a chi-squared test. However, as discussed above, the average treatment effect would be of limited value if policyholders show significantly heterogeneity in response to the marketing intervention activity. Our objective is to estimate the personalized treatment effect and use it to construct an optimal treatment rule for the Auto Insurance portfolio – i.e., the policyholder-treatment assignment that maximizes the expected profits from the campaign.

Table 2: Cross-sell rates by group

	Treatment	Control
Purchased Home policy = N	30,184	3,322
Purchased Home policy = Y	789	75
Cross-sell rate	2.55%	2.21%

Note. This table displays the cross-sell rate for the treatment and control groups. The average treatment effect is **0.34%** (2.55% - 2.21%), which is not statistically significant (P value = 0.23).

To objectively examine the performance of the proposed method, we randomly split the data into training and validation sets in a 70/30 ratio. A preliminary analysis showed that model performance is not highly sensitive to the values of its tuning parameters (i.e., number of trees B and number of variables n randomly sampled as candidates at each split), as long as they are specified within a reasonable range. Thus, we fitted a causal conditional inference forest (`ccif`) to the training data using its default parameter values. Specifically, in Algorithm 2, we used $B = 500$, $n = 16$, and a P value = 0.05 as the level of significance α. We next ranked policyholders in the validation data set based on their estimated personalized treatment effect (from high to low), and grouped them into deciles. We then computed the actual average treatment effect within each decile (defined as the difference in cross-sell rates between the treatment and control groups).

Figure 3 shows the boxplots of the actual average treatment effect for each decile based on 100 random training/validation data partitions. The results show that clients with higher estimated personalized treatment effect were, on average, positively influenced to buy as a result of the marketing intervention activity. Also, notice there is a subgroup of clients whose purchase behaviour was negatively impacted by the campaign. Negative reactions to sales attempts has been recognized in the literature (Günes et al., 2010; Kamura, 2008; Byers and So, 2007) and may happen for a variety of reasons. For instance, the marketing activity may trigger a decision to shop for better multi-product rates among other insurers. Moreover, if the client currently owns a Home policy with another insurer, she may decide to switch the Auto policy to that insurer instead. We found evidence of higher Auto policy cancellation rates at the higher deciles. In addition, some clients may perceive the call as intrusive and likely be annoyed by it, generating a negative reaction.

In the context of insurance, it is not only important to consider the personalized treatment effect from the cross-sell activity, but the risk profile of the targeted clients (Thuring et al., 2012; Kaishev et al., 2013; Englund et al., 2009). After taking into account the expected life-time-value of a Home policy[6] and the fixed and variable expenses from the campaign, we determined the expected profitability from targeting each decile. Based on these considerations, Figure 3 shows that only clients in deciles 1-3 have positive expected profits from the marketing activity and should be targeted. The incremental profits from clients in deciles 4-7 is outweighed by the incremental costs, and so the company should avoid targeting these clients. Clients in deciles 8-10 have negative reactions to the campaign and clearly should not be targeted either.

7. Conclusions

The estimation of personalized treatment effects is becoming increasingly important in many scientific disciplines and policy making. As subjects can show significant heterogeneity in response to treatments, making an optimal treatment choice at the individual subject level is essential. An

[6]The expected life-time-value (LTV) of a Home policy in decile $i = \{1, \ldots, 10\}$ is given by $LTV_i = [\bar{P}_i - \hat{LC}_i - \overline{EXP}_i] \sum_{t=1}^{5} \text{Prob}(S_{it}) r^t$, where $\bar{P}$ is the average policy premium, $\hat{LC}$ is the predicted insurance losses per policy-year, $\overline{EXP}$ captures the fixed and variable expenses for servicing the policy, $\text{Prob}(S_{it})$ is the probability that a policyholder in decile $i = \{1, \ldots, 10\}$ will survive with the Home product beyond year $t = \{1, \ldots, 5\}$, and r^t is the interest discount factor.

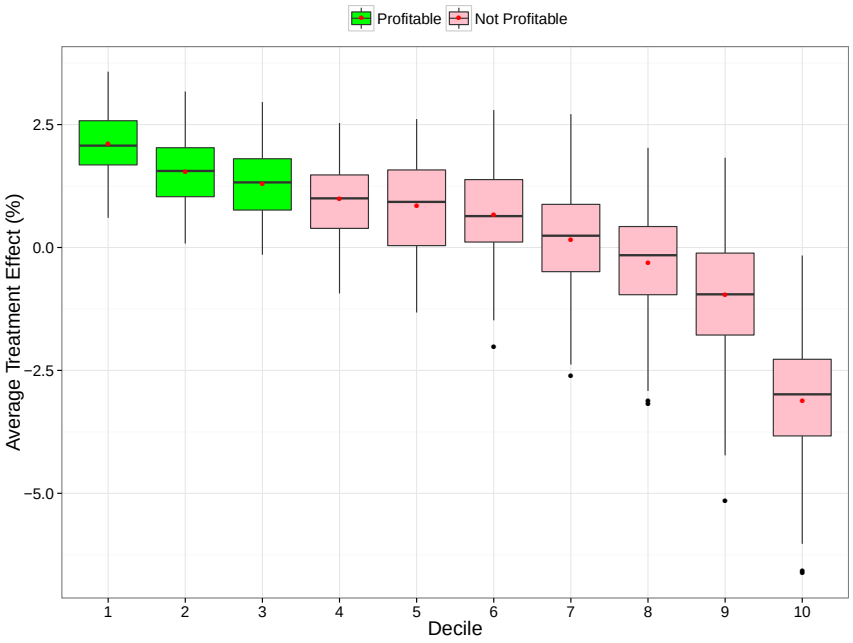

Figure 3: *Boxplots of the actual average treatment effect for each decile based on 100 random training/validation data splits. The first (tenth) decile represents the 10% of clients with highest (lowest) predicted personalized treatment effect. Clients with higher estimated personalized treatment effect were, on average, positively influenced to buy* as a result *of the marketing intervention activity.*

optimal personalized treatment is the one that maximizes the probability of a desirable outcome. We call the task of learning the optimal personalized treatment *personalized treatment learning*.

From the statistical learning perspective, estimating personalized treatment effects imposes some key challenges, primarily because the optimal treatment is unknown on a given training set. In this paper, we discussed seven of the most prominent methods proposed in the literature to tackle this problem, and proposed a new approach called *causal conditional inference trees*. Our method recursively partitions the input space into subgroups with heterogeneous treatment effects. Motivated by the *unbiased recursive partitioning* method proposed by Hothorn et al. (2006), the key ingredient of our tree-based method is the separation between the variable selection and the splitting procedure, coupled with a statistically motivated and computational efficient stopping criteria based on the theory of permutation tests developed by Strasser and Weber (1999). This statistical approach prevents overfitting, without requiring any form of pruning or cross-validation. It also avoids selection bias towards covariates with many possible splits. Performance results measured on synthetic data show that our proposed method often outperforms the alternatives on the numerical settings described in this article.

We have also discussed an application of the proposed method in the context of insurance marketing for the purpose of selecting the best targets for cross-selling an insurance product. Our method was able to identify the policyholders who were positively/negatively motivated to buy as a result of the marketing intervention activity. Based on marketing costs considerations, we next derived the policyholder-treatment assignment that maximizes the expected profitability from the campaign.

We would also like to acknowledge the limitations of this work. First, we have only considered the case of binary treatments. It would be worthwhile to examine the extent to which the methods discussed in this article can be extended to multi-category or continuos treatment settings. Second, in many situations the interest may be to estimate the personalized treatment effect when the intervention is not applied on a randomized basis, but we think there are major background variables that influence which treatment is received. Thus, it would be relevant to consider personalized treatment learning models in the context of observational data. Finally, we have only consider

the case of personalized treatments in a single-decision setup. In dynamic treatment regimes, the treatment type is repeatedly adjusted according to an ongoing individual response (Murphy, 2005). In this context, the goal is to optimize a set of time-varying personalized treatments for the purpose of maximizing the probability of a long-term desirable outcome.

Acknowledgements

LG thanks Royal Bank of Canada, RBC Insurance. MG and AMP-M thanks ICREA Academia and the Ministry of Science / FEDER grant ECO2010-21787-C03-01.

References

Abu-Mostafa, Y., Magdon-Ismail, M. and Hsuan-Tien, L. 2012. *Learning From Data.* AMLBook.

Alemi, F., Erdman, H., Griva, I. and Evans, C. 2009. Improved statistical methods are needed to advance personalized medicine. *Open Transl Med J.* 1: 16–20.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 57(1): 289–300.

Brieman, L., Friedman, J., Olshen, R. and Stone, C. 1984. *Classification and Regression Trees.* New York: Chapman & Hall.

Brieman, L. 2001. Statistical modeling: the two cultures. *Statistical Science* 16(3): 199–231.

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

Byers, R. and So, K. 2007. Note - A mathematical model for evaluating cross-sales policies in telephone service centers. *Manufacturing & Service Operations Management* 9(1): 1–8.

Chawla, N. 2005. Data mining for imbalanced datasets: An overview. *Data Mining and Knowledge Discovery Handbook.*, Springer US.

Cover, T. and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1): 21–27.

Cover, T. and Thomas, J. 1991. *Elements of Information Theory*, Second Edition. John Wiley & Sons, Inc.

Dawid, A. 1979. Conditional independence in statistical theory. *Journal of the Royal Statistical Society B* 41(1): 1–31.

Dehejia, R. and Wahba, S. 1999. Causal effects in non experimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94: 1053–1062.

Englund, M., Gustafsson, J., Nielsen, J. and Thuring, F. 2009. Multidimensional credibility with time effects: An application to commercial business lines. *Journal of Risk and Insurance* 76(2): 443–453.

Estabrooks, A., Jo, T. and Japkowicz, N. 2004. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence* 20(1): 18–36.

Frawley, W., Piatetsky-Shapiro, G. and Matheus, C. 1991. Knowledge discovery in databases – An overview. *Knowledge Discovery in Databases*: 1–30.

Friedman, J. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38: 367–378.

Guelman, L., Guillén, M. and Pérez-Marín, A.M. 2012. Random forests for uplift modeling: an insurance customer retention case. *Lecture Notes in Business Information Processing* 115: 123–133.

Guelman, L., Guillén, M. and Pérez-Marín, A.M. 2013. Uplift random forests. *Cybernetics & Systems* , *forthcoming.*

Guelman, L. and Guillén, M. 2014. A causal inference approach to measure price elasticity in automobile insurance. *Expert Systems with Applications* 41: 387–396.

Günes, E., Aksin-Karaesmen, O., Örmeci, L. and Özden, H. 2010. Modeling customer reactions to sales attempts: If cross-selling backfires. *Journal of Service Research* 13(2): 168–183.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning*, Second Edition. New York: Springer.

Holland, P. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81(396): 945–960.

Holland, P. and Rubin, D. 1988. Causal inference in retrospective studies. *Evaluation Review* 12: 203–231.

Hothorn, T., Hornik, K. and Zeileis, A. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3): 651–674.

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

Jaśkowski, M. and Jaroszewicz, S. 2012. Uplift modeling for clinical trial data. *ICML 2012 Workshop on Clinical Data Analysis*, Edinburgh, Scotland, UK, 2012.

Kaishev, V., Nielsen, J. and Thuring, F. 2013. Optimal customer selection for cross-selling of financial services products. *Expert Systems with Applications* 40(5): 1748–1757.

Kamakura, W. 2008. Cross-selling: Offering the right product to the right customer at the right time. *Journal of Relationship Marketing* 6(3-4): 41–58.

Kass, G. 1980. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 29(2): 119–127.

LaLonde, R. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76(4): 606–620.

Larsen, K. 2009. Net models. *M2009* - 12^{th} *Annual SAS Data Mining Conference.*

Liang, H., Xue, Y. and Berger, B. 2006. Web-based intervention support system for health promotion. *Decision Support Systems* 42(1): 435–449.

Lo, V. 2002. The true lift model. *ACM SIGKDD Explorations Newsletter* 4(2): 78–86.

Murphy, S. 2005. An experimental design for the development of adaptive treatment strategies. *Statist. Med.* 24:

1455–1481

Qian, M. and Murphy, S. 2011. Performance guarantees for individualized treatment rules. *Annals of Statistics* 39(2) 1180–1210.

Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Radcliffe, N. and Surry, P. 2011. Real-World Uplift Modelling with Significance-Based Uplift Trees. *Portrait Technical Report* TR-2011-1.

Rosenbaum, P. and Rubin, D. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5): 688–701.

Rubin, D. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 2: 1–26.

Rubin, D. 1978. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* 6: 34–58.

Rubin, D. 2005. Causal inference using potential outcomes. *Journal of the American Statistical Association* 100(469): 322–330.

Rubin, D. and Waterman, R. 2006. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science* 21: 206–222.

Rzepakowski, P. and Jaroszewicz, S. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems* 32(2): 303–327

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

Sinha, A. and Zhao, H. 2008. Incorporating domain knowledge into data mining classifiers: An application in indirect lending. *Decision Support Systems* 46(1): 287–299.

Strasser, H. and Weber, C. 1999. On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics* 8: 220–250.

Su, X., Tsai, C., Wang, H., Nickerson, D. and Li, B. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* 10(2): 141–158.

Su, X., Kang, J., Fan, J., Levine, R. and Yan, X. 2012. Facilitating score and causal inference trees for large observational studies. *Journal of Machine Learning Research* 13(10): 2955–2994.

Tang, H., Liao, S. and Sun, S. 2013. A prediction framework based on contextual data to support mobile personalized marketing. *Decision Support Systems*, In Press.

Thuring, F., Nielsen, J., Guillén, M. and Bolancé, C. 2012. Selecting prospects for cross-selling financial products using multivariate credibility. *Expert Systems with Applications* 39(10): 8809–8816.

Tian, L., Alizadeh, A., Gentles, A. and Tibshirani, R. 2012. A simple method for detecting interactions between a treatment and a large number of covariates. *Submitted on Dec 2012*. arXiv:1212.2995v1 [stat.ME].

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

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Wahba, G. 2002. Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences* 99(26): 16524–16530.

Weiss, G. and Provost, F. 2003. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research* 19: 315–354.

Wright, P. 1992. Adjusted p-values for simultaneous inference. *Biometrics* 48: 1005–1013.

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

Xu, D., Liao, S. and Li, Q.. 2008. Combining empirical experimentation and modeling techniques: A design research approach for personalized mobile advertising applications. *Decision Support Systems* 44(3): 710–724.

Zhao, Y. and Zeng, D. 2012. Recent development on statistical methods for personalized medicine discovery. *Frontiers of Medicine* 7(1): 102–110.

Žliobaitė, I. and Pechenizkiy, M. 2010. Learning with actionable attributes: Attention – boundary cases! *ICDMW '10 Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*: 1021–1028.

Appendix

Maximum likelihood estimates of personalized treatment effects from the Modified Covariate and Modified Outcome methods are equivalent.

From the modified outcome method, we have under the logistic model for binary response

$$E\{l(W, g(\mathbf{X}))|\mathbf{X}, A = 1\} = E(W|\mathbf{X} = \mathbf{x}, A = 1)g(\mathbf{X}) - \log(1 + e^{g(\mathbf{X})}),$$

and

$$E\{l(W, g(\mathbf{X}))|\mathbf{X}, A = 0\} = E(W|\mathbf{X} = \mathbf{x}, A = 0)g(\mathbf{X}) - \log(1 + e^{g(\mathbf{X})}),$$

where $g(\mathbf{X}) = \beta^\top \mathbf{X}$.

Thus,

$$
\begin{aligned}
\mathcal{L}(g) &= E\{l(W, g(\mathbf{X}))\} \\
&= E_{\mathbf{X}}\Big[\frac{1}{2}E_W\{l(W, g(\mathbf{X}))|\mathbf{X}, A=1\} + \frac{1}{2}E_W\{l(W, g(\mathbf{X}))|\mathbf{X}, A=0\}\Big] \\
&= E_{\mathbf{X}}\Big[\frac{1}{2}\{E(Y|\mathbf{X}, A=1)g(\mathbf{X}) - \log(1+e^{g(\mathbf{X})})\} + \\
&\quad \frac{1}{2}\{(1-E(Y|\mathbf{X}, A=0))g(\mathbf{X}) - \log(1+e^{g(\mathbf{X})})\}\Big] \\
&= \frac{1}{2}E_{\mathbf{X}}\Big[\tau(\mathbf{X})g(\mathbf{X}) + g(\mathbf{X}) - 2\log(1+e^{g(\mathbf{X})})\Big],
\end{aligned}
$$

where $\tau(\mathbf{X}) = E[Y|\mathbf{X}=\mathbf{x}, A=1] - E[Y|\mathbf{X}=\mathbf{x}, A=0]$.

Therefore,

$$
\frac{\partial \mathcal{L}}{\partial g} = \frac{1}{2}E_{\mathbf{X}}\Big[\tau(\mathbf{X}) + 1 - 2\frac{e^{g(\mathbf{X})}}{(1+e^{g(\mathbf{X})})}\Big].
$$

Thus,

$$
g^*(\mathbf{x}) = \log\Big\{\frac{1+\tau(\mathbf{x})}{1-\tau(\mathbf{x})}\Big\},
$$

or equivalently,

$$
\tau(\mathbf{x}) = \frac{e^{g^*(\mathbf{x})}-1}{e^{g^*(\mathbf{x})}+1}.
$$

That is, the loss minimizer of $\mathcal{L}(g)$, $g^*(\mathbf{x})$, is equal to $f^*(\mathbf{x})$ in (7), which is the loss minimizer of $E\{Yf(\mathbf{X})A - \log(1+\mathbf{exp}(f(\mathbf{X})A))\}$ from the modified covariate method.