Causal Inference via Sparse Additive Models with Application to Online Advertising

Wei Sun[1], **Pengyuan Wang**[2], **Dawei Yin**[2], **Jian Yang**[2], **and Yi Chang**[2]
[1]Purdue University, West Lafayette, IN,USA, sun244@purdue.edu
[2]Yahoo Labs, Sunnyvale, CA, USA, {pengyuan, daweiy, jianyang,yichang}@yahoo-inc.com

Abstract

Advertising effectiveness measurement is a fundamental problem in online advertising. Various causal inference methods have been employed to measure the causal effects of ad treatments. However, existing methods mainly focus on linear logistic regression for univariate and binary treatments and are not well suited for complex ad treatments of multi-dimensions, where each dimension could be discrete or continuous. In this paper we propose a novel two-stage causal inference framework for assessing the impact of complex ad treatments. In the first stage, we estimate the propensity parameter via a sparse additive model; in the second stage, a propensity-adjusted regression model is applied for measuring the treatment effect. Our approach is shown to provide an unbiased estimation of the ad effectiveness under regularity conditions. To demonstrate the efficacy of our approach, we apply it to a real online advertising campaign to evaluate the impact of three ad treatments: ad frequency, ad channel, and ad size. We show that the ad frequency usually has a treatment effect cap when ads are showing on mobile device. In addition, the strategies for choosing best ad size are completely different for mobile ads and online ads.

Introduction

In the current online advertising ecosystem, user are exposed to ads with diverse formats and channels, and user's behaviors are caused by complex ad treatments combining of various factors (Rosenkrans 2009). The complexity of ad treatments calls for an accurate and causal measurement of ad effectiveness, i.e., how the ad treatment *causes* the changes in outcomes. An outcome is the user's response in the ad campaigns, such as whether or not the user clicks a link, or searches for a brand. The gold standard of accurate ad effectiveness measurement is the experiment-based approach, such as A/B test, where different ad treatments are randomly assigned to users. However, the cost of fully randomized experiments is usually very high (Chan et al. 2010; Kohav and Longbotham 2010; Stitelman et al. 2011) and in some rich ad treatment circumstances, such fully randomized experiments are even infeasible (Bottou et al. 2013). Hence it is critical and necessary to provide statistical approaches to estimate the ad effectiveness directly from observational data rather than experimental data. Previous studies based on observational data try to establish direct relationship between the ad treatment and a success action (Abraham 2008). However, in observational data, typically the user characteristics may affect both the exposed ad treatment and the success tendency. Ignoring such confounding effects of user characteristics may lead to a biased estimation of the treatment effect (Rosenbaum and Rubin 1984).

Causal inference aims to infer unbiased causality effect of the ad treatment from observational data, by eliminating the impact of the confounding factors such as user characteristics. The causal inference framework for a binary treatment was original proposed by Rosenbaum and Rubin (1983). They introduced the concept of propensity score and proposed to estimate it via a logistic regression. To expand the scope of causal inference from a binary treatment to a continuous/categorical treatment, a propensity function-based framework was proposed (Imai and van Dyk 2004), where Gaussian linear regression was applied to estimate the propensity function. Based on the estimated propensity score (or propensity function), an additional sub-classification procedure was usually performed to produce the final treatment effect estimation. Recently, machine learning algorithms, e.g. gradient boosted machine, bagged CART, and random forest, were employed to estimate the propensity score (Lee, Lessler, and Stuart 2010; Su et al. 2012). While these algorithms showed desirable performance, they mainly focused on the scenario with a univariate treatment and low dimensional user features. In online advertising, causal inference methods have been developed to estimate the causal effect (Chan et al. 2010; Dalessandro et al. 2012; Wang, Traskin, and Small 2013; Wang et al. 2014). Nevertheless, these works also mainly focused on the univariate and binary treatment and not well suited for measuring the effects of complex treatment (e.g., a two-dimensional treatment consists of ad exposure frequency and ad size), which are more realistic and in great demand.

Measuring complex ad treatment effectiveness is still facing four major challenges. First, the general ad treatment can be much more complex than binary treatment. It could be a discrete or continuous, uni- or multi-dimensional treatment.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Second, the widely enforced linearity condition on the relationship between ad treatment and user characteristics is too restricted. Third, the online dataset typically has high-dimensional user characteristics and it is generally believed that only a small number of features are truly informative while others are redundant. Forth, existing propensity-based sub-classification methods are sensitive to the choice of the number of sub-classes and the format of sub-classification.

To address above challenges, in this paper we propose a novel two-stage causal inference framework to tackle the general ad effectiveness measurement problem. In the first stage, we model the propensity parameter via a sparse additive model and in the second stage, we employ a propensity-adjusted regression model to measure the final treatment effect. The novelty and advantages of the proposed method can be summarized as follows:

- Our causal inference is fully general, where the treatment can be single- or multi-dimensional, and it can be binary, categorical, continuous, or a mixture of them. We prove that this framework offers an unbiased estimation of the treatment effect under standard assumptions.
- Our sparse additive model for propensity parameter estimation deals with high-dimensionality and non-linearity issues in online advertising data.
- The propensity-adjusted regression model in our second stage estimates the treatment effect directly and avoids the tuning in existing sub-classification methods.

We further apply our framework to an online advertising campaign and provide practical guideline to assess advertising strategy on mobile and online platforms. Our extensive experiments show that the ad frequency usually has a treatment effect cap on mobile devices and the choice of best ad size are completely different for ads shown on mobile and online. Hence it is important for the ad providers to make appropriate adjustment for the number of the ads and the format of the ads delivered to the users.

Background

We first review the basic concepts of causal inference, and then mention potential drawbacks of existing approaches.

Define a treatment as a random variable $\mathbf{T}$ and a potential outcome associated with a specific treatment $\mathbf{T} = \mathbf{t}$ as $Y(\mathbf{t})$. Since the treatment can be uni- or multi-dimensional, we use the boldface $\mathbf{T}$ and $\mathbf{t}$ to indicate a multivariate treatment variable and T and t to indicate a univariate treatment variable. In general, multivariate variable $\mathbf{T}$ could be of a mixture of categorical and continuous variables. For each user, indexed by $i = 1, 2, ..., N$, we observe a vector of pretreatment covariates (i.e., user characteristics) $\mathbf{X}_i$ of length p, a treatment $\mathbf{T}_i$, and an univariate outcome Y_i corresponding to the treatment received. Typically, one would like to evaluate the effect of a given treatment $\mathbf{t}$ on the outcome Y, removing the confounding effect of $\mathbf{X}$.

A primary interest of causal inference is the distribution $p(Y(\mathbf{t}))$ for each treatment $\mathbf{t}$. In order to unbiasedly evaluate this distribution, two standard assumptions are usually made in the literature (Rosenbaum and Rubin 1983).

Assumption 1: Stable unit treatment value assumption. The potential outcome for one unit should be unaffected by the particular assignment of treatments to the other units.

Assumption 2: Strong ignorability of treatment assignment. Given the covariates $\mathbf{X}$, the distribution of treatment $\mathbf{T}$ is independent of the potential outcome $Y(\mathbf{t})$ for all $\mathbf{t}$.

Assumption 1 allows us to model the outcome of one subject independent of another subject's treatment status, given the covariates. Assumption 2 enables the modeling of the treatment with respect to the covariates, independent of the outcome, i.e., all the features related to both the treatment assignment and the outcome have been included in the model. Under Assumption 2, Rosenbaum and Rubin (1983) showed that the distribution $p(Y(\mathbf{t}))$ can be computed as

$$p(Y(\mathbf{t})) = \int_{\mathbf{X}} p(Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, \mathbf{X}) p(\mathbf{X}) d\mathbf{X}, \qquad (1)$$

where $p(Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, \mathbf{X})$ is the conditional distribution of $Y(\mathbf{t})$ given $\mathbf{t}$ and $\mathbf{X}$, and $p(\mathbf{X})$ is the distribution of $\mathbf{X}$. In order to compute (1), one can model $p(Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, \mathbf{X})$ directly. But experience shows that the result can be strongly biased if the relationship between $\mathbf{T}$ and $\mathbf{X}$ is omitted or misspecified (Rosenbaum and Rubin 1983). When the observed covariates $\mathbf{X}$ is low-dimensional, one way to avoid this bias is to classify subjects according to $\mathbf{X}$ and estimate (1) via the weighted average over $\mathbf{X}$. However as the dimension of $\mathbf{X}$ increases, exact sub-classification according to covariates becomes computationally infeasible.

To address these issues, Rosenbaum and Rubin (1983) introduced the balancing score to summarize the information required to balance the distribution of covariates and proposed the propensity score method for the binary treatment problem. The balancing score is the random variable such that conditioned on it, the observed covariates and the treatment assignment are independent. Later on, Imai and van Dyk (2004) generalized propensity score to propensity function for a categorical or continuous treatment. Specifically, the propensity function $e(\mathbf{X})$ is defined as the conditional density of the treatment given the observed covariates, i.e., $e(\mathbf{X}) = p(\mathbf{T}|\mathbf{X})$. It was shown that this propensity function is a balancing score, that is, $p(\mathbf{T}|\mathbf{X}) = p(\mathbf{T}|e(\mathbf{X}))$. Hence we can obtain $p(Y(\mathbf{t}))$ in (1) as

$$p(Y(\mathbf{t})) = \int_{e(\mathbf{X})} p(Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, e(\mathbf{X})) p(e(\mathbf{X})) de(\mathbf{X}). \quad (2)$$

To compute the integral in (2), Imai and van Dyk (2004) assumed that there existed a unique finite-dimensional propensity parameter $\boldsymbol{\theta}$ such that the propensity function $e(\mathbf{X})$ depended on $\mathbf{X}$ only through $\boldsymbol{\theta}(\mathbf{X})$. Here $\boldsymbol{\theta}$ is also a balancing score, and hence we can obtain $p(Y(\mathbf{t}))$ in (2) as

$$p(Y(\mathbf{t})) = \int_{\boldsymbol{\theta}} p(Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \qquad (3)$$

Usually $\boldsymbol{\theta}$ has a much smaller dimension than X, hence this strategy tackles the high dimensionality issue of the covariates in (1).

In many advertising applications, $e(\mathbf{X})$ or $\boldsymbol{\theta}$ is unknown since the collected data are observational data. Although the

ad publisher designs the advertising algorithm, the specific $e(\mathbf{X})$ may not be fully controlled due to the system complexity. More importantly, the probability of a specific user visiting an ad publisher is unknown, which is an important factor of $e(\mathbf{X})$. For example, some ads are only posted on given pages, and it is unknown whether a specific user visits this page or not. Therefore, in practice, we need to estimate $\boldsymbol{\theta}$ based on given samples.

To approximate (3), Imai and van Dyk (2004) suggested to estimate the propensity parameter $\boldsymbol{\theta}$ via Gaussian linear regression (for a continuous treatment) or linear logistic regression (for a binary treatment), classify samples into several sub-classes with similar value of the estimations of $\boldsymbol{\theta}$, estimate the treatment effect within each sub-class, and then average the estimators from each sub-class.

However, linear models for estimating $\boldsymbol{\theta}$ require restricted assumptions on functional form and distributions of variables. The violation of these assumptions may result in a biased treatment effect estimation (Lee, Lessler, and Stuart 2010). In addition, the final treatment effect estimation could be sensitive to the number of sub-classes and the strategy of sub-classification (Hullsiek and Louis 2002). Note that a larger number of sub-classes leads to a more accurate estimation of the integral in (3) but inevitably implies a less accurate estimation of the inner conditional distribution due to limited observations in each sub-class. Furthermore, although equal-frequency strategy is generally used to form the sub-classes (Rosenbaum and Rubin 1984; Imai and van Dyk 2004), experiments showed that this strategy often leads to highly unstable estimators for the extreme sub-classes (Hullsiek and Louis 2002). Therefore, it is necessary to introduce a new method which has a more flexible modeling strategy and can avoid the specifications of the number of sub-classes and the format of sub-classification.

Methodology

To address the above drawbacks, we propose a two-stage causal inference method for more realistic problem—complex ad effectiveness analysis. The method applies a more flexible modeling approach for estimating the propensity parameter and avoids specifying the number of sub-classes. The outline of our algorithm is shown in Table 1 and Figure 1. The detailed modeling approaches in Stages 1-2 are discussed in the following subsections.

Table 1: Our Two-stage Algorithm

Input:	Y_i, $\mathbf{X}_i$, $\mathbf{T}_i$ for $i = 1, 2, ...N$.
Output:	Estimated treatment effect for $\mathbf{t}$.
Stage 1:	Obtain the estimated propensity parameter $\widehat{\boldsymbol{\theta}}(\mathbf{X}_i)$ by modeling $\mathbf{T}_i \vert \mathbf{X}_i$ via SAM.
Stage 2:	Calculate the final treatment effect by modeling $Y_i \vert \mathbf{T}_i, \widehat{\boldsymbol{\theta}}(\mathbf{X}_i)$ via GAM as in (10).

Stage 1: Propensity Parameter Estimation via SAM

We propose to estimate the propensity parameter via a sparse additive model (SAM) which relieves the restricted linear assumption in existing propensity parameter estimation approaches, and deals with the high-dimensionality issue of $\mathbf{X}$ where only a small number of them are truly informative. For a m-dimensional treatment $\mathbf{T} = (T_1, \ldots, T_m)$, we estimate the propensity parameter $\boldsymbol{\theta}_j$, $j = 1, \ldots, m$, by modeling each $T_j|\mathbf{X}$ via SAM. The final estimated propensity parameter for $\mathbf{T}$ is the combination of all the individually estimated propensity parameters, i.e., the tuple $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$. In this sequel, we will discuss the modeling approach for a continuous treatment, a binary treatment, and a multi-class categorical treatment, separately. For ease of notation, we will suppress the dependence of T_j on the index j in this subsection and consider a generic modeling for $T|\mathbf{X}$ for a univariate treatment T.

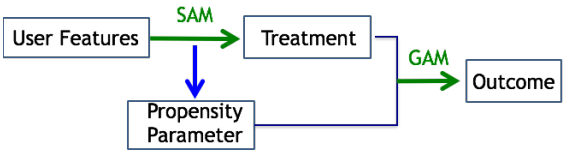

Figure 1: Outline of our algorithm. "SAM" and "GAM" refer to sparse additive model and generalized additive model, respectively.

Continuous treatment: For a continuous treatment T, Imai and van Dyk (2004) estimated the propensity parameter via a Gaussian linear regression $T = \mathbf{X}^T\beta + \epsilon$ with $\mathbf{X}$ the p-dimensional user features and ϵ a standard Gaussian. Despite of its simplicity, the linearity assumption is too restrictive. To relax this constraint, Woo et. al. (2008) applied the generalized additive model (GAM) (Hastie and Tibshirani 1990) to estimate the propensity parameter. For smooth functions $f_1, \ldots, f_p$, the additive model assumes $T = \sum_{j=1}^{p} f_j(X_j) + \epsilon$, which is more general than Gaussian linear regression, but only has good statistical and computational performance when p is relatively small.

In high-dimensional scenario, the sparsity constraint has been incorporated into the additive model for variable selection. The key idea of SAM (Ravikumar et al. 2009) is to scale each component function by a scalar β_j and then impose an l_1 penalty on the coefficient $\beta = (\beta_1, \ldots, \beta_p)^T$ to encourage the sparsity of the nonlinear components. Denote $\mathcal{H}_j$ as the Hilbert space of measurable functions $f_j(x_j)$ of the single variable x_j with $\mathbb{E}[f_j(X_j)] = 0$ and $\mathbb{E}[f_j^2(X_j)] < \infty$. The SAM solves

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} \quad \mathbb{E}\left(T - \sum_{j=1}^{p} \beta_j g_j(X_j)\right)^2 \tag{4}$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \le \lambda \tag{5}$$

$$\mathbb{E}(g_j) = 0, j = 1 \ldots, p \tag{6}$$

$$\mathbb{E}(g_j^2) = 1, j = 1 \ldots, p. \tag{7}$$

Similar to the lasso penalty (Tibshirani 1996) for Gaussian linear regression, the upper bound constraint in (5) encourages the sparsity, which results in a sparse additive function $\sum_{j=1}^{p} f_j(x_j) = \sum_{j=1}^{p} \beta_j g_j(x_j)$. Furthermore, the last two

constraints (6) and (7) are used for identifiability. The above optimization problem can be solved efficiently by the coordinate descent algorithm (Ravikumar et al. 2009), where the tuning parameter λ can be tuned by BIC or stability-based criterion (Sun, Wang, and Fang 2013). After solving (4), we can estimate the propensity parameter $\boldsymbol{\theta}(\mathbf{X})$ using the vector of the fitted treatments $\widehat{T}$.

Binary treatment: The SAM in (4) can be extended to nonparametric logistic regression for estimating the propensity parameter $\boldsymbol{\theta}(\mathbf{X}) = \mathbb{P}(T = 1|\mathbf{X})$. Specifically, for a binary treatment $T \in \{0, 1\}$, the additive logistic model is

$$\mathbb{P}(T = 1|\mathbf{X}) = \frac{\exp\left(\sum_{j=1}^{p} f_j(X_j)\right)}{1 + \exp\left(\sum_{j=1}^{p} f_j(X_j)\right)}$$

with the population log-likelihood function $L(f) = \mathbb{E}[Tf(\mathbf{X}) - \log(1 + \exp(f(\mathbf{X})))]$. Then sparse additive logistic model can be solved by replacing the squared loss in (4) with the corresponding new log-likelihood $L(f)$. The propensity parameter $\boldsymbol{\theta}(\mathbf{X})$ for a binary treatment can then be estimated as the vector of fitted conditional probabilities $\widehat{\mathbb{P}}(T = 1|X_i)$ with $i = 1 \ldots, N$.

Multi-class treatment: When the treatment T is categorical with more than 2 classes, e.g., $T \in \{1, \ldots, K\}$, the one-vs.-rest multi-category strategy (Liu, Lafferty, and Wasserman 2008) can be employed. Specifically, the treatment T is expanded as a $K-1$ dimensional vector $(T^{(1)}, \ldots, T^{(K-1)})$ in which at most one element can be one and all others being zero. The multi-category additive logistic model is,

$$\mathbb{P}(T^{(k)} = 1|\mathbf{X}) = \frac{\exp\left(\sum_{j=1}^{p} f_j^{(k)}(X_j)\right)}{1 + \sum_{k'=1}^{K-1} \exp\left(\sum_{j=1}^{p} f_j^{(k')}(X_j)\right)},$$

for $k = 1, \ldots, K-1$ with the multinomial log-loss

$$l(f) = \sum_{k=1}^{K-1} T^{(k)} f^{(k)}(\mathbf{X}) - \log\left(1 + \sum_{k=1}^{K-1} \exp(f^{(k)}(\mathbf{X}))\right).$$

Finally, the propensity parameter for a K-class categorical treatment can be estimated as the $N \times (K-1)$ matrix whose (i, k)-th element is the estimated conditional probability $\widehat{\mathbb{P}}(T = k|X_i)$ for $k = 1, \ldots, K-1$ and $i = 1 \ldots, N$.

Stage 2: Propensity-adjusted Regression

With the propensity parameter $\boldsymbol{\theta}$, we now measure the treatment effect via the propensity-adjusted regression, which can more naturally estimate the treatment effect than the existing sub-classification methods. In short, we consider the propensity parameter as an additional feature and directly model the relationship between the outcome and the combination of treatment and propensity parameter via regression models. Specifically, given the propensity parameter $\boldsymbol{\theta}$ and the treatment $\mathbf{T}$, we model $Y|\mathbf{T}, \boldsymbol{\theta}$ as

$$Y = g\big(\mathbf{T}, \boldsymbol{\theta}\big) + \epsilon \tag{8}$$

with $g(\cdot)$ an unknown function and ϵ a standard Gaussian.

Next we show that the treatment effect estimator based on (8) is unbiased under some regularity conditions.

Theorem 1 *Under Assumptions 1-2, and assume true outcome model is* (8)*, if there exists an unbiased functional estimator $\widehat{g}$ for g, i.e., $\mathbb{E}_{\mathcal{D}}[\widehat{g}] = g$ where the expectation is with respect to the samples $\mathcal{D}$, then the estimated treatment effect of $\mathbf{t}$ is unbiased. That is,*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{\mathcal{D}}[\widehat{g}(\boldsymbol{t}, \boldsymbol{\theta})] - \mathbb{E}_{\mathcal{D}}[\widehat{g}(\boldsymbol{\theta}, \boldsymbol{\theta})]\right] = \mathbb{E}[Y(\boldsymbol{t}) - Y(\boldsymbol{\theta})].^{1} \tag{9}$$

Proof of Theorem 1: First, for any given treatment $\mathbf{t}$, we have $\mathbb{E}[Y(\mathbf{t})] = \mathbb{E}_{\boldsymbol{\theta}}\{\mathbb{E}[Y(\mathbf{t})|\boldsymbol{\theta}]\}$. Since the propensity parameter $\boldsymbol{\theta}$ is a balancing score, according to Theorem 3 in Rosenbaum and Rubin (1983) and Assumption 2, we have

$$(Y(\mathbf{t}), Y(\mathbf{0})) \perp\!\!\!\perp \mathbf{T}|\boldsymbol{\theta}.$$

Therefore, we have $\mathbb{E}[Y(\mathbf{t})|\boldsymbol{\theta}] = \mathbb{E}[Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, \boldsymbol{\theta}]$ for any treatment $\mathbf{t}$. According to (8), we have $\mathbb{E}[Y(\mathbf{t})|\mathbf{T} = \mathbf{t}, \boldsymbol{\theta}] = g(\mathbf{t}, \boldsymbol{\theta})$ and $\mathbb{E}[Y(\mathbf{0})|\mathbf{T} = \mathbf{0}, \boldsymbol{\theta}] = g(\mathbf{0}, \boldsymbol{\theta})$. Combining these results with the unbiasedness of $\widehat{g}$ leads to (9). This ends the proof of Theorem 1. ■

We next discuss an estimation procedure of the treatment effect based on (8). Given the training samples $\{(Y_i, \mathbf{T}_i, \mathbf{X}_i,), i = 1, \ldots, n\}$, in Stage 1, we obtain the estimated propensity parameter $\widehat{\boldsymbol{\theta}}$ via SAM, and in Stage 2 we estimate the function $\widehat{g}(\cdot)$ in (8) via GAM, where GAM is applied since it offers flexible modeling and performs well when dimensions of $\mathbf{T}$ and $\boldsymbol{\theta}$ are low. Finally, the averaged treatment effect (ATE) of a treatment $\mathbf{t}$ can be estimated as

$$\widehat{\mathrm{ATE}(\mathbf{t})} = \frac{1}{N} \sum_{i=1}^{N} \left\{\widehat{g}(\mathbf{t}, \widehat{\boldsymbol{\theta}}(X_i)) - \widehat{g}(\mathbf{0}, \widehat{\boldsymbol{\theta}}(X_i))\right\}. \tag{10}$$

Simulations

This section evaluates the efficacy of our model via a simulated experiment. We first illustrate the performance of model fitting in Stage 1 by comparing our approach with up-to-date machine learning algorithms, and then demonstrate the superiority of our treatment effect estimation.

We set sample size $N = 1000$ and number of features $p = 200$. Define basis functions $f_1(x) = -2\sin(2x)$, $f_2(x) = x^2 - 1/3$, $f_3(x) = x - 0.5$, $f_4(x) = e^{-x} - e^{-1} - 1$, $f_5(x) = (x - 0.5)^2 + 2$, $f_6(x) = \mathbb{I}_{\{x>0\}}$, $f_7(x) = e^{-x}$, $f_8(x) = \cos(x)$, $f_9(x) = x^2$, and $f_{10}(x) = x$. We first generate the features $X_1, \ldots, X_p \overset{iid}{\sim} N(0, 1)$. The three dimensional treatment $\mathbf{T}$ and the outcome Y are produced as follows. We generate the continuous treatment $T_1|X \sim N(\sum_{j=1}^{4} f_j(X_j), 1)$, the binary treatment $T_2|X = 1$ if $\sum_{j=1}^{5} f_j(X_j) > 0$ and 0 otherwise, and the multi-class treatment $T_3|X \in \{1, 2, 3, 4\}$ based on 25%, 50%, 75%, and 100% quantiles of $\sum_{j=1}^{2} f_j(X_j)$, and generate the outcome model as $Y|X, \mathbf{T} \sim N(\sum_{j=1}^{5} f_{j+5}(X_j) + \alpha^T \mathbf{T}, 1)$ with $\alpha = (1, 1, 1)$.

In this example, the features $X_1, \ldots, X_5$ are informative variables and have confounding effects to both treatment and outcome, and the rest $X_6, \ldots, X_p$ are noisy variables to mimic the sparse data of online advertising. The true effects for three treatment T_1, T_2, or T_3 are all 1.

[1] We implicitly consider $\mathbf{t} = \mathbf{0}$ as the baseline treatment.

To evaluate the performance of model fitting in Stage 1, we independently generate a testing data with size 1000, and report the averaged prediction errors for fitting T_1 and the averaged misclassification errors for fitting T_2 over 100 replications. The results for fitting T_3 are similar and are omitted to save space. We compare SAM with gradient boosted machine (gbm), lasso, sparse logistic regression (slogit), bagged tree (bagging), and random forest (rf). As shown in Figure 2, SAM achieves smallest errors for both continuous and binary treatment model fittings and the advantages over other models are significant.

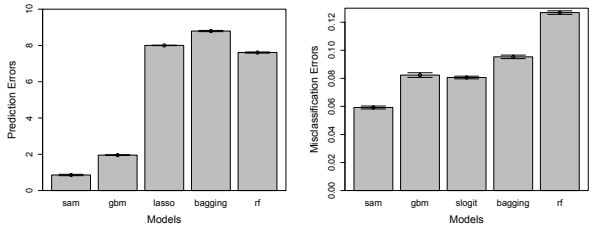

Figure 2: Errors of fitting T_1 (left plot) and T_2 (right plot). "sam" refers to sparse additive model, "gbm" refers to gradient boosted machine,"slogit" refers to sparse logistic regression, "bagging" refers to bagged tree, and "rf" refers to random forest. The one standard deviation bar is shown for each model.

The benefit of SAM is not at the cost of expensive computations. As shown in Figure 3, SAM is the second fastest algorithm when fitting the continuous treatment, which is only slightly slower than lasso, and its computational cost is comparable to gbm when fitting the binary treatment.

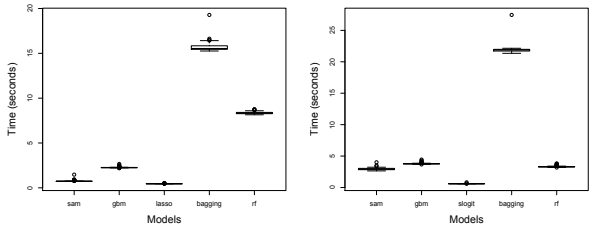

Figure 3: Computational time (in seconds) for fitting treatment T_1 (left plot) and treatment T_2 (right plot) from various models.

Secondly, we demonstrate the superior treatment effect estimation performance of our model. Specifically, we compute the absolute difference of the estimated treatment effects and their corresponding true treatment effects. We investigate the comparison among a direct treatment effect estimation method (Model 1), the existing propensity-adjusted methods (Models 2-3) as well as their sparse versions (Models 4-5), and the proposed two-stage models.

Model 1: a linear regression for $Y|(\mathbf{T}, \mathbf{X})$.

Model 2: fit a linear regression for $T_1|\mathbf{X}$, logistic regression for $T_2|\mathbf{X}$, multi-class logistic regression for $T_3|\mathbf{X}$ to obtain the propensity parameter matrix $\widehat{\boldsymbol{\theta}}$ and then fit a linear regression for $Y|\mathbf{T}, \widehat{\boldsymbol{\theta}}$.

Model 3: same as **Model 2** except that Stage 2 fits a GAM for $Y|\mathbf{T}, \widehat{\boldsymbol{\theta}}$.

Model 4: same as **Model 2** except that Stage 1 fits a linear sparse model for $T_j|\mathbf{X}$, $j = 1, 2, 3$.

Model 5: same as **Model 3** except that Stage 1 fits a linear sparse model for $T_j|\mathbf{X}$, $j = 1, 2, 3$.

Our Model 1: fit a SAM for $T_j|\mathbf{X}$, $j = 1, 2, 3$ to obtain $\widehat{\boldsymbol{\theta}}$ and fit a linear regression for $Y|\mathbf{T}, \widehat{\boldsymbol{\theta}}$.

Our Model 2: same as **Our Model 1** except that Stage 2 fits a GAM for $Y|\mathbf{T}, \widehat{\boldsymbol{\theta}}$.

In Figure 4, we report the sum of errors of all three treatment effect estimations over 100 replications. Clearly, our models achieve the smallest errors and their advantages are significant. For our two models, fitting a GAM in Stage 2 provides with more accuracy gain. Besides, comparing Model 2, Model 4, and our Model 1 reveals that the accuracy of treatment fitting in Stage 1 is extremely important for the final treatment effect estimation. Furthermore, the sparse models (Models 4-5) outperform their non-sparse counterparts (Models 2-3), which ensures the importance of feature selection in the high dimensional scenario.

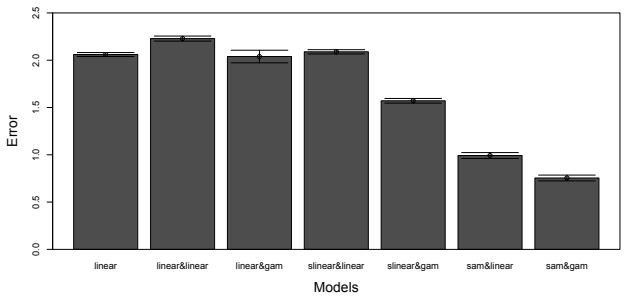

Figure 4: Errors of treatment effect estimations. "linear", "linear&linear", "linear&gam", "slinear&linear", "slinear&gam" refer to Models 1-5, and "sam&linear", "sam&gam" are our Models 1-2.

Experiments

We apply our model on a real advertising campaign for a major auto company from a premium internet media company. This campaign involves advertisements delivered via mobile devices (including iPhone, iPad, Android Phone, Android Tablet, Windows Phone and Windows Tablet) and personal computers (PCs). We aim to measure the exposure frequency impact and the size impact of the ads from both platforms.

The datasets contains about 5 millions of users, among which 1.6 thousand users perform success actions (online quotes) [2]. The advertising treatment **T** is a 3-dimensional

[2]The reported dataset and results are deliberately incomplete

vector: The first dimension is the ad frequency which can be treated as a continuous variable; the second dimension is the device indicator which is a binary variable with 1 referring to a mobile and 0 referring to a PC; and the third dimension is the ad size indicator which is a categorical variable with 3 classes: small, medium, and large. We consider ad frequency of values at most 10 which consists of 99% samples. In total, there are 2483 features, including the demographic information, personal interest, and online and TV activities. The personal interest includes high-dimensional and sparse features, and hence SAM is well suited for the modeling.

We first estimate the propensity parameter via SAM and compare it with estimators via gradient boosted machine and sparse linear model, where the latter two demonstrate good empirical performance in the simulations. The accuracy of the propensity parameter estimation is evaluated via the prediction errors by refitting the model for each treatment. The averaged prediction errors over all the treatments are summarized in Figure 5. Clearly, SAM achieves minimal errors and is significantly better than other two models.

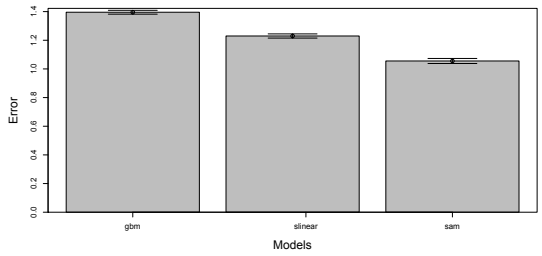

Figure 5: Prediction errors of treatment fitting in real example. The"gbm", "slinear", and "sam" refer to gradient boosted machine, sparse linear model, and sparse additive model, respectively.

We then illustrate the findings on mobile and desktop platforms from three perspectives: 1) the ad frequency impact ; 2) the ad size impact; and 3) the synthetic impact of ad frequency and ad size.

The ad frequency impact for both mobile devices and PCs. As shown in Figure 6, when the customers are exposed to ads via the mobile devices, the number of conversions monotonically increases as the ad exposure increases. When customers are exposed to ads via the PCs, the number of conversions increases at the beginning and then decreases, and the maximal number of conversions is obtained when the users are shown 7 ads via PCs. These results advise that 7 ads are sufficient to maximize the number of conversions for online advertisements and there is little demand for this auto company to deliver more than 7 ads to the eligible users; while they should continue to show ads to customers who are exposed to less than 10 mobile ads. In addition, Figure 6 delivers that the ad frequency effects on mobile devices are marginally much larger than that on PCs. Specifically, exposing the user to 5 ads via mobile, as opposed to 1 ad via mobile will lead to about 5 extra conversions; while exposing the user to 5 ads via PCs, as opposed to 1 ad via PCs will lead to only 1 extra conversion. This suggests that in this campaign users generally has a larger chance to convert when they are shown mobile ads, which is consistent with the observation in Butcher (2010) that mobile ad campaigns are generally more effective than online norms.

and subject to anonymization, and thus do not necessarily reflect the real portfolio at any particular time.

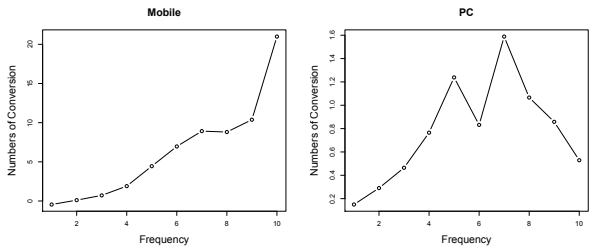

Figure 6: Fitted number of conversions versus the ad frequency for mobile devices (left) and PCs (right). The X axis is the ad frequency and Y axis is the fitted number of conversions based on our model.

The ad size impact for mobile devices and PCs from our causal inference estimation in Figure 7. A general trend is that, when the ads are shown via mobile devices, the smaller ad delivers more conversions than the larger ad. In contrast, when the ads are displayed on PCs, larger ad size triggers more conversions. This advises that on PCs it is more economic to split the large ads into several small ads to achieve more conversions.

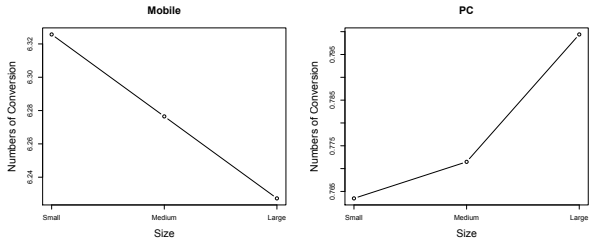

Figure 7: Fitted number of conversions versus the ad size for mobile devices (left) and PCs (right). The X axis is the ad size and Y axis is the fitted number of conversions based on our model.

The synthetic impact of ad frequency and ad size. We group ad frequencies as 1-2, 3-4, 5-6, 7-8, and 9-10 buckets for easy presentation. As illustrated in Figure 8, when the ads are shown on mobile devices, the largest conversion is obtained when the users are shown $9-10$ ads with small ad size; On the other hand, the largest conversion is obtained when the users are shown $7-8$ ads with large ad size via the PCs. Therefore, it is crucial for the ad providers to make appropriate adjustment based on the number and size of the ads the users have been exposed to.

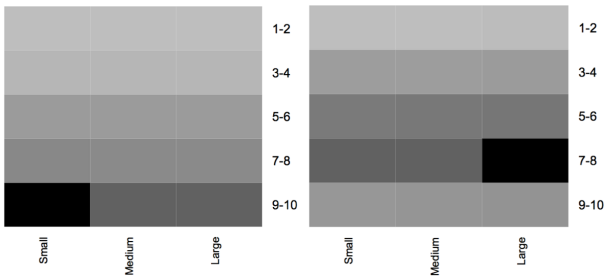

Figure 8: Heatmaps of fitted number of conversions for mobile devices (left plot) and PCs (right plot). The rows are the ad frequencies and the columns are the ad sizes.

Conclusion

In this paper we propose a novel two-stage causal inference framework for assessing the impact of complex advertising treatments. In Stage 1, we utilize SAM for propensity parameter estimation, which is essentially a non-parametric method which is more general and flexible than linear models and is suited for sparse advertising data with complex treatments. In Stage 2, we devise a propensity-adjusted regression to estimate treatment effect, which can more naturally estimate the treatment effect than the existing subclassification methods. Our model is theoretically unbiased and outperforms existing approaches in extensive experiments. Our approach is applied to a real campaign with both mobile and online ads to investigate the impact of ad frequency, ad size, and the synthetic impact across platforms. Our approach successfully draws meaningful insights from the complex dataset and provides practical guideline to the advertisers.

Note that our framework is not limited to online advertising, but is also applicable to other user engagement studies where causal impact of general treatments (e.g., UI design, content format, ad context) needs to be measured with observational data.

References

Abraham, M. 2008. The off-line impact of online ads. *Harvard Business Review* 86(4): 28.

Bottou, L.; Peters, J.; Quinonero-Candela, J.; Charles, D. X.; Chickering, D. M.; Portugaly, E.; Ray, D.; Simard, P.; and Snelson, E. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research* 14:3207–3260.

Butcher, D. 2010. *Mobile ad campaigns 5 times more effective than online: InsightExpress study*. http://www.mobilemarketer.com.

Chan, D.; Ge, R.; Gershony, O.; Hesterberg, T.; and Lambert, D. 2010. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of SIGKDD*, 7–16. ACM.

Dalessandro, B.; Perlich, C.; Stitelman, O.; and Provost, F. 2012. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, 7. ACM.

Hastie, T. J., and Tibshirani, R. J. 1990. *Generalized Additive Models*. Chapman and Hall/CRC.

Hullsiek, K., and Louis, T. 2002. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* 2:179–193.

Imai, K., and van Dyk, D. A. 2004. Causal inference with general treatment regimes. *Journal of the American Statistical Association* 99:854–866.

Kohav, R., and Longbotham, R. 2010. Unexpected results in online controlled experiments. *SIGKDD Explorations* 12(2).

Lee, B.; Lessler, J.; and Stuart, E. 2010. Improving propensity score weighting using machine learning. *Statistics in Medicine* 29:337–346.

Liu, H.; Lafferty, J.; and Wasserman, L. 2008. Nonparametric regression and classification with joint sparsity constraints. In *Proceedings of NIPS*. ACM.

Ravikumar, P.; Lafferty, J.; Liu, H.; and Wasserman, L. 2009. Sparse additive models. *Journal of the Royal Statistical Society, Series B* 71:1009–1030.

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

Rosenbaum, P. R., and Rubin, D. B. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79(387):516–524.

Rosenkrans, G. 2009. The creativeness & effectiveness of online interactive rich media advertising. *Journal of Interactive Advertising* 9(2):18–31.

Stitelman, O.; Dalessandro, B.; Perlich, C.; and Provost, F. 2011. Estimating the effect of online display advertising on browser conversion. *Data Mining and Audience Intelligence for Advertising* 8.

Su, X.; Kang, J.; Fan, J.; Levine, R. A.; and Yan, X. 2012. Facilitating score and causal inference trees for large observational studies. *The Journal of Machine Learning Research* 13(1):2955–2994.

Sun, W.; Wang, J.; and Fang, Y. 2013. Consistent selection of tuning parameters via variable selection stability. *The Journal of Machine Learning Research* 14:3419–3440.

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

Wang, P.; Liu, Y.; Meytlis, M.; Tsao, H.-Y.; Yang, J.; and Huang, P. 2014. An efficient framework for online advertising effectiveness measurement and comparison. In *Proceedings of WSDM*. ACM.

Wang, P.; Traskin, M.; and Small, D. S. 2013. Robust inferences from a before-and-after study with multiple unaffected control groups. *Journal of Causal Inference* 1–26.

Woo, M.; Reiter, J.; and Karr, A. 2008. Estimation of propensity scores using generalized additive models. *Statistics in Medicine* 27:3805–3816.