The Causal Learning of Retail Delinquency

Yiyan Huang,[2*] Cheuk Hang Leung,[2*] Xing Yan,[3] Qi Wu,[2†]
Nanbo Peng,[1] Dongdong Wang,[1] Zhixiang Huang[1]
[1]JD Digits
[2]City University of Hong Kong
[3]ISBD, Renmin University of China
yiyhuang3-c@my.cityu.edu.hk, {chleung87, qiwu55}@cityu.edu.hk, xingyan@ruc.edu.cn
{pengnanbo, wangdongdong9, huangzhixiang}@jd.com

Abstract

This paper focuses on the expected difference in borrower's repayment when there is a change in the lender's credit decisions. Classical estimators overlook the confounding effects and hence the estimation error can be magnificent. As such, we propose another approach to construct the estimators such that the error can be greatly reduced. The proposed estimators are shown to be unbiased, consistent, and robust through a combination of theoretical analysis and numerical testing. Moreover, we compare the power of estimating the causal quantities between the classical estimators and the proposed estimators. The comparison is tested across a wide range of models, including linear regression models, tree-based models, and neural network-based models, under different simulated datasets that exhibit different levels of causality, different degrees of nonlinearity, and different distributional properties. Most importantly, we apply our approaches to a large observational dataset provided by a global technology firm that operates in both the e-commerce and the lending business. We find that the relative reduction of estimation error is strikingly substantial if the causal effects are accounted for correctly.

Introduction

A growing number of technology conglomerates provide lending services to shoppers who frequent their e-commerce marketplaces. Technology firms have the information advantage that commercial banks lack. A vast amount of proprietary digital footprints are now at their fingertips. Study shows the information content of proxies for human behavior, lifestyle, and living standard in the non-transnational data are just effective, hence highly valuable for default prediction (Berg et al. 2020). However, managing retail credit risks in online marketplaces differs in one fundamental way from managing credit-card default risks faced by commercial banks. It comes from the pronounced "action-response" relationship between the lender's credit decisions and the borrowers' delinquency outcomes. When customers apply for credit to finance their online purchases, they effectively enter into an unsecured loan contract where the counterparty is the platform lender, and they are expected to make installments according to the payment schedules. These loan decisions, especially the loan amount and loan interest, are far more individualized than those made in the traditional credit card business and are frequently adjusted. The e-commerce lenders observe that their credit policies have a systematic causal impact, which alters borrowers' payment behavior, irrespective of age, income, education, occupation, and so forth.

It is conceivable that machine learning (ML) algorithms can effectively tap into the information advantage for accurate estimations of delinquency rate (Khandani, Kim, and Lo 2010). However, the existing retail credit risk studies do not recognize the essential of action-response causalities as far as we know. Indeed, obtaining an accurate estimation of the delinquency rate does not mean that we can obtain an accurate estimation of the action-response causality. Generally, we face two biases in estimating the action-response causality. The first bias is due to the neglect of the confounding effects. In reality, loan decisions are the results of the lender's decision algorithms that use an overlapping set of borrowers' features with those used for risk assessments, e.g., past shopping and financing records. Thus, ignoring the modelling of the relation between the credit policies and the credit features may cause a big bias in estimating the action-response causality. The second bias comes from the estimation of the predictors-response relation. This is a regression bias related to data and ML algorithm, e.g., the sample size, the feature dimensions and the regressor selections.

The goal of this paper is to construct the estimators of the causal parameters which can address both the confounding bias and the regression bias. It is equivalent to finding the score functions with specific conditions (detailed discussions will be presented in the later sections) such that we can recover the estimators from the score functions. To obtain the corresponding estimates of the estimators, we need to estimate the "counterfactuals", which are the potential outcomes in delinquent probabilities of borrowers if they were given different amounts of credit lines from the ones they had received. Once the counterfactuals are estimated, we would like to assess both the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATTE). For new customers, the lender can use the ATE to

[*] Co-first authors are in alphabetical order.

[†] Qi Wu is the corresponding author.

choose appropriate credit lines for them. For existing customers, the ATTE allows the lender to gauge the potential changes of risks if their credit lines were changed from the current levels to new ones.

For the rest of the paper, Section "Related Works and Contributions" summarizes the related works and our contributions. Section "The Model Setup" presents our model setup, and Section "Experiments" presents all the experiments. The paper ends with Section "Conclusion".

Related Works and Contributions

Related Works. In the past credit risk works, the typical supervised learning methods such as direct regression are widely used to estimate the global relationship between the response and the predictors. For example, the regression tree (Khandani, Kim, and Lo 2010), Random Forests (Malekipirbazari and Aksakalli 2015), and Recurrent Neural Network (Sirignano, Sadhwani, and Giesecke 2016) are applied to construct default forecasting models.

However, when it comes to studying the action-response relationship (e.g., estimating the ATE and ATTE), simply using the typical supervised learning methods to estimate the outcomes for each individual under different interventions and averaging the estimated outcomes for each intervention can produce unsatisfactory results (Chernozhukov et al. 2018), since there is a chance of misspecification of the relationship when confounding effects are present. Some other methods that can account for the confounding effects range from those based on balancing learning and matching (Li and Fu 2017; Kallus 2018; Bennett and Kallus 2019) to those based on deep learning such as representation learning and adversarial learning (Johansson, Shalit, and Sontag 2016; Yao et al. 2018; Lattimore, Lattimore, and Reid 2016; Yoon, Jordon, and van der Schaar 2018). However, for matching methods such as Propensity Score Matching (PSM) and inverse probability weighting (IPW) (e.g., (Hirano, Imbens, and Ridder 2003)), they may amplify the estimation bias if the feature variables are not selected properly or the algorithm is not ideal enough (Heckman, Ichimura, and Todd 1998).

To overcome some deficiencies of the above methods, Doubly Robust Estimators (DREs) are proposed (e.g., (Farrell 2015)). The DREs are recovered from the score functions which incorporate the confounding effects in general (Dudík, Langford, and Li 2011). However, it is not sure if the score functions of the DREs satisfy the *orthogonal condition*, which is defined in (Chernozhukov et al. 2018) inspired by (Neyman 1979). Heuristically, those DREs recovered from the score functions which may violate the orthogonal condition (see the detailed discussions in our supplementary) can be sensitive to the nuisance parameters, and hence easily lead to a biased estimation. To solve this problem, some researchers propose the new score functions that satisfy the orthogonal condition (Chernozhukov et al. 2018; Mackey, Syrgkanis, and Zadik 2018; Oprescu, Syrgkanis, and Wu 2019), but they either only consider the binary treatment or derive the theoretical results based on the partially linear model (PLR) setting. To improve that, we not only extend the intervention variable from the binary values to the multiple values, but also consider the fully nonlinear model setting instead of the PLR model setting.

Contributions. The contributions of our paper are:

1. We are the first to show the importance and necessity of considering causality in retail credit risk study using observational records of e-commerce lenders and borrows;
2. We extend from a partially linear model (e.g., (Mackey, Syrgkanis, and Zadik 2018; Oprescu, Syrgkanis, and Wu 2019)) to a fully nonlinear model. Our estimators recovered from the orthogonal score functions are proved to be regularization unbiased and consistent with an increasing number of population size, thus very suitable for large datasets. Besides, our setup allows the intervention variable to take multiple discrete values, rather than binary values as in (Chernozhukov et al. 2018);
3. Our estimators are generic, not only restricted to linear or tree-based methods, but also for any complex methods such as neural network-based models including fully-connected ones (e.g., MLP), convolutional ones (e.g., CNN), and recurrent ones (e.g., GRU);
4. The obtained estimators are robust to model misspecifications when mappings between predictors and outcomes/interventions exhibit different degrees of nonlinearity, and data possess different distributional properties than assumed;
5. We use the parameters in our experiments to control the causality and nonlinearity and show how much error our estimators can correct for compared with direct regression estimators used in the past credit risk works.

The above points are comprehensively tested through well-designed simulation experiments and verified on a large proprietary real-world dataset via semi-synthetic experiments.

The Model Setup

The Potential Outcomes. Given a probability space $(\Omega, \mathbb{P}, \mathscr{F})$, the formulation (1) treats the treatment variable D (or the policy intervention) as part of the explanatory variables $(D, \mathbf{Z})$, where $\mathbf{Z}$ is the feature set. D and $\mathbf{Z}$ are used to regress the response variable (or the outcome) Y such that

$$Y = g(D, \mathbf{Z}) + \zeta \quad \text{and} \quad \mathbb{E}\left[\zeta \mid D, \mathbf{Z}\right] = 0. \tag{1}$$

Here the outcome Y is a random scalar, the feature set $\mathbf{Z}$ is a random vector, and the intervention D takes discrete values in the set $\{d^1, \cdots, d^n\}$. ζ is the scalar noise which summarizes all other non-$(D, \mathbf{Z})$-related unknown factors and is assumed to have zero conditional mean given $(D, \mathbf{Z})$. The function g is a $\mathbb{P}$-integrable function. The core of the impact inference is the estimation of the potential outcomes of an individual under the interventions that are different from what we observe. The unobservable potential outcomes are also called the *counterfactuals*. Knowing how to estimate the counterfactuals provides a way to estimate two quantities, both of them are the average "action-response" relationship between the potential outcome Y and the potential treatment d^i. Mathematically, they are

$$\theta^i := \mathbb{E}\left[g(d^i, \mathbf{Z})\right] \quad \text{and} \quad \theta^{i|j} := \mathbb{E}\left[g(d^i, \mathbf{Z}) \mid D = d^j\right].$$

The quantity θ^i means that we want to find the expected outcome when the potential intervention is d^i. Concurrently, the quantity $\theta^{i|j}$ means that given the factual intervention is d^j, we want to find the expected counterfactual outcome if the intervention D had taken the value d^i.

Impact Inference without Confounding. We begin with the Impact Inference without Confounding (IoC) which accounts for the policy impact but not the confounding effects. We use $(y_m, d_m, \mathbf{z}_m)$ to represent the observational data associated with the m^{th} customer in the size-N population and use $(y_m^j, d_m^j, \mathbf{z}_m^j)$ to represent the observational data of the m^{th} customer in the sub-population which contains N_j customers with observed treatment d^j. Using sample averaging, we can estimate θ^i as $\frac{1}{N}\sum\limits_{m=1}^{N} g(d^i, \mathbf{z}_m)$.

On the other hand, since

$$\begin{aligned}&\mathbb{E}\left[g(d^i,\mathbf{Z})\mathbf{1}_{\{D=d^j\}}\right] = \mathbb{E}\left[\mathbb{E}\left[g(d^i,\mathbf{Z})\mathbf{1}_{\{D=d^j\}} \mid D\right]\right]\\ =&\mathbb{E}\left[g(d^i,\mathbf{Z}) \mid D=d^j\right]\mathbb{E}\left[\mathbf{1}_{\{D=d^j\}}\right],\end{aligned}$$

we can obtain that $\theta^{i|j} = \frac{\mathbb{E}\left[g(d^i,\mathbf{Z})\mathbf{1}_{\{D=d^j\}}\right]}{\mathbb{E}\left[\mathbf{1}_{\{D=d^j\}}\right]}$. Simultaneously, we use $\frac{1}{N}\sum\limits_{m=1}^{N}\mathbf{1}_{\{d_m=d^j\}} g(d^i,\mathbf{z}_m)$ to estimate $\mathbb{E}\left[g(d^i,\mathbf{Z})\mathbf{1}_{\{D=d^j\}}\right]$. Consequently, we can estimate $\theta^{i|j}$ as $\frac{1}{N_j}\sum\limits_{m=1}^{N}\mathbf{1}_{\{d_m=d^j\}} g(d^i,\mathbf{z}_m)$.

Note that θ^i is an average over the whole population N, whereas $\theta^{i|j}$ is an average over the sub-population N_j. To compute the estimates of the two quantities, the key point is to obtain $\hat{g}$ (the estimate of g) for every d^i from the observational dataset. In the related literature, the specification of g includes both the additive forms (Djebbari and Smith 2008) and the multiplicative forms (Hainmueller, Mummolo, and Xu 2019). The choices of $\hat{g}$ also include linear models (Du and Zhang 2015; Li and Bell 2017) as well as nonlinear ones. In particular, the predictive advantage of using neural networks to estimate g is demonstrated in (Shi, Blei, and Veitch 2019), where (Louizos et al. 2017; Yoon, Jordon, and van der Schaar 2018) formulate a similar relationship to estimate the treatment effects. Other examples related to the estimations of the treatment effects include (Alaa and Schaar 2018; Toulis and Parkes 2016; Li and Fu 2017; Syrgkanis et al. 2019).

Once we obtain $\hat{g}(d^i,\cdot)$ (the estimate of $g(d^i,\cdot)$), θ^i and $\theta^{i|j}$ can be respectively estimated in the IoC context as

$$\hat{\theta}_o^i = \frac{1}{N}\sum_{m=1}^{N}\hat{g}(d^i,\mathbf{z}_m) \quad \text{and} \quad \hat{\theta}_o^{i|j} = \frac{1}{N_j}\sum_{m=1}^{N_j}\hat{g}(d^i,\mathbf{z}_m^j).$$

We call them the IoC estimates since we omit the relationship between D and $\mathbf{Z}$, or the so-called confounding effects. Indeed, the confounding effects are obscured from the IoC estimates.

Impact Inference with Confounding. We then propose the Impact Inference with Confounding (IwC) which can account for both the policy impact and the confounding effects. Using the IoC estimates to estimate θ^i and $\theta^{i|j}$ can be misspecified (Dudík, Langford, and Li 2011; Yuan et al. 2019) when the confounding effects are present. The misspecification comes from the fact that, while $\mathbf{Z}$ in (1) affects the outcome Y, the intervention D could also be driven by the confounding variable $\mathbf{Z}$. For example, the customer's income level could be a confounding variable in the retail lending context. Customers who receive higher credit lines usually have higher incomes, and higher-income people tend to have lower credit risk. Such examples also widely exist in recommendation systems (e.g., (Wang et al. 2019; Swaminathan and Joachims 2015a,b) and the references therein). In our paper, we construct better score functions satisfying the orthogonal condition such that the corresponding estimates are assured to be regularization unbiased (Chernozhukov et al. 2018).

In order to study the policy impact in the presence of confounding effects, we propose the following formulation for our IwC estimations:

$$Y = g(D,\mathbf{U},\mathbf{Z}) + \xi, \quad \mathbb{E}[\xi \mid \mathbf{U},\mathbf{X},\mathbf{Z}] = 0, \tag{2a}$$

$$D = m(\mathbf{X},\mathbf{Z},\nu), \qquad \mathbb{E}[\nu \mid \mathbf{X},\mathbf{Z}] = 0, \tag{2b}$$

where $\mathbf{Z}$ is the confounder, $\mathbf{U}$ is the outcome-specific feature set, and $\mathbf{X}$ is the intervention-specific feature set. The map m and the noise term ν are of the same nature as those of g and ξ. We impose no functional restrictions on g and m such that they can be parametric or non-parametric, linear or nonlinear, etc.

If the confounding effects (2b) are not recognized but present in the data, the IoC estimates

$$\hat{\theta}_o^i = \frac{1}{N}\sum_{m=1}^{N}\hat{g}(d^i,\mathbf{u}_m,\mathbf{z}_m),\ \hat{\theta}_o^{i|j} = \frac{1}{N_j}\sum_{m=1}^{N_j}\hat{g}(d^i,\mathbf{u}_m^j,\mathbf{z}_m^j) \tag{3}$$

could be inaccurate. Indeed, we ignore the impact caused by (2b) when (3) are used. Thus, the estimated outcome-predictors relation $\hat{g}$ can have opposite outcome-predictors relation of the authentic g w.r.t. the features variable. Furthermore, even if $\hat{g}$ has the similar relation with g w.r.t. the features variable, (3) can be *regularization biased*, meaning that the estimators are sensitive to the estimation of $\hat{g}$. As such, we should construct the regularization unbiased estimators which use the information given in (2a) and (2b). PSM, IPW and doubly robust estimators (DREs) (e.g., (Farrell 2015)) are methodologies which incorporate the relation (2b) through the computation of *propensity score*. However, these approaches can be sensitive w.r.t. the small perturbations on the map m in (2b). Consequently, it may not be suitable for the estimations of ATE and ATTE. To stabilize it, we should build the estimators which can be recovered from the score functions that satisfy the orthogonal condition in the Definition 1. Heuristically, the partial derivative of score functions w.r.t. the nuisance parameters are expected to be 0. Indeed, the estimation errors of the ATE and ATTE are reduced using a multiplicative term of the propensity score and the residuals between the observed Y and the estimate of $g(d^i,\cdot,\cdot)$.

Definition 1 (Orthogonal Condition). *Let W be the random elements, Θ be a convex set which contains the causal parameter ϑ of dimension d_ϑ (θ is the true causal parameter we are interested in) and T be a convex set which contains the nuisance parameter ϱ (ρ is the true nuisance parameter we are interested). Moreover, we define the Gateaux derivative map $D_{r,j}[\varrho - \rho] := \partial_r \left\{\mathbb{E}[\psi_j(W, \theta, \rho + r(\varrho - \rho))]\right\}$. We say that a score function ψ satisfies the (Neyman) orthogonal condition if for all $r \in [0, 1)$, $\varrho \in \mathscr{T} \subset T$ and $j = 1, \cdots, d_\vartheta$, we have*

$$\partial_\varrho \mathbb{E}[\psi_j(W, \theta, \varrho)] \mid_{\varrho=\rho} [\varrho - \rho] := D_{0,j}[\varrho - \rho] = 0. \quad (4)$$

To start with our IwC formulation, we let $W = (Y, D, \mathbf{X}, \mathbf{U}, \mathbf{Z})$. The quantities Θ, T, ϑ, θ, ϱ and ρ stated in the Definition 1 for the score function of θ^i are

$$\begin{aligned}
&\Theta = \Theta_i := \{\vartheta = \mathbb{E}\left[\mathscr{g}(d^i, \mathbf{U}, \mathbf{Z})\right] \mid \mathscr{g} \textit{ is } \mathbb{P}\textit{-integrable}\},\\
&T = T_i := \{\varrho = \left(\mathscr{g}(d^i, \mathbf{U}, \mathbf{Z}), a_i(\mathbf{X}, \mathbf{Z})\right) \mid \mathscr{g} \textit{ is } \mathbb{P}\textit{-integrable}\},\\
&\theta = \theta^i := \mathbb{E}\left[g(d^i, \mathbf{U}, \mathbf{Z})\right] \in \Theta_i,\\
&\rho = \rho^i := \left(g(d^i, \mathbf{U}, \mathbf{Z}), \mathbb{E}\left[\mathbf{1}_{\{D=d^i\}} | \mathbf{X}, \mathbf{Z}\right]\right) \in T_i.
\end{aligned}$$

Similarly, the quantities Θ, T, ϑ, θ, ϱ and ρ stated in the Definition 1 for the score function of $\theta^{i|j}$ are

$$\begin{aligned}
&\Theta = \Theta_{i|j}\\
&\quad := \{\vartheta = \mathbb{E}\left[\mathscr{g}(d^i, \mathbf{U}, \mathbf{Z}) \mid D = d^j\right] \mid \mathscr{g} \text{ is } \mathbb{P}\text{-integrable}\},\\
&T = T_{i|j} := \Big\{\varrho = (\mathscr{g}(d^i, \mathbf{U}, \mathbf{Z}), m_j,\\
&\qquad\qquad a_j(\mathbf{X}, \mathbf{Z}), a_i(\mathbf{X}, \mathbf{Z})) \mid \mathscr{g} \text{ is } \mathbb{P}\text{-integrable}\right\},\\
&\theta = \theta^{i|j} := \mathbb{E}\left[g(d^i, \mathbf{U}, \mathbf{Z}) \mid D = d^j\right] \in \Theta_{i|j},\\
&\rho = \rho^{i|j} := \Big(g(d^i, \mathbf{U}, \mathbf{Z}), \mathbb{E}\left[\mathbf{1}_{\{D=d^j\}}\right],\\
&\qquad\qquad \mathbb{E}\left[\mathbf{1}_{\{D=d^j\}} | \mathbf{X}, \mathbf{Z}\right], \mathbb{E}\left[\mathbf{1}_{\{D=d^i\}} | \mathbf{X}, \mathbf{Z}\right]\Big) \in T_{i|j}.
\end{aligned}$$

Here, $\mathscr{g}(d^i, \mathbf{U}, \mathbf{Z})$, $a_i(\mathbf{X}, \mathbf{Z})$ and m_j are the arbitrary nuisance parameters, while $g(d^i, \mathbf{U}, \mathbf{Z})$, $\mathbb{E}\left[\mathbf{1}_{\{D=d^i\}} \mid \mathbf{X}, \mathbf{Z}\right]$ and $\mathbb{E}\left[\mathbf{1}_{\{D=d^j\}}\right]$ are the corresponding true nuisance parameters we are interested in. Our aim is to construct the score functions such that the moments of the Gateaux derivative of the score functions w.r.t. $\mathscr{g}(d^i, \mathbf{U}, \mathbf{Z})$, $a_i(\mathbf{X}, \mathbf{Z})$ and m_j evaluated at $g(d^i, \mathbf{U}, \mathbf{Z})$, $\mathbb{E}\left[\mathbf{1}_{\{D=d^i\}} \mid \mathbf{X}, \mathbf{Z}\right]$ and $\mathbb{E}\left[\mathbf{1}_{\{D=d^j\}}\right]$ are 0, implying the Definition 1 holds.

Before stating the score functions, we introduce some notations to simplify our expressions. We define the estimate of $g(d^i, \mathbf{x}, \mathbf{z})$ as $\hat{g}(d^i, \mathbf{x}, \mathbf{z})$. Furthermore, we define $P_i(\mathbf{x}, \mathbf{z}) = \mathbb{E}[\mathbf{1}_{\{D=d^i\}} \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}]$ and the corresponding estimate as $\hat{P}_i(\mathbf{x}, \mathbf{z})$ for any $i = 1, \cdots, n$. To find $\hat{P}_i(\mathbf{x}, \mathbf{z})$, we can use any classification methods to obtain it. For example, when we use Logistic regression to estimate $\hat{P}_i(\mathbf{x}, \mathbf{z})$, it becomes $1/\left[1 + \exp\left(-\mathbf{w}_x^T \mathbf{x} - \mathbf{w}_z^T \mathbf{z} - \mathbf{w}\right)\right]$.

Theorem 2. *The score function $\psi^i(W, \vartheta, \varrho)$ which can be used to recover an estimate of θ^i and satisfies the Definition 1 is*

$$\vartheta - \mathscr{g}(d^i, \mathbf{U}, \mathbf{Z}) - \frac{\mathbf{1}_{\{D=d^i\}}}{a_i(\mathbf{X}, \mathbf{Z})}(Y - \mathscr{g}(d^i, \mathbf{U}, \mathbf{Z})), \quad (5)$$

while the score function $\psi^{i|j}(W, \vartheta, \varrho)$ which can be used to recover an estimate of $\theta^{i|j}$ and satisfies the Definition 1 is

$$\begin{aligned}
\frac{1}{m_j}\Big\{&\vartheta \mathbf{1}_{\{D=d^j\}} - \mathscr{g}(d^i, \mathbf{U}, \mathbf{Z})\mathbf{1}_{\{D=d^j\}}\\
&-\mathbf{1}_{\{D=d^i\}}\frac{a_j(\mathbf{X}, \mathbf{Z})}{a_i(\mathbf{X}, \mathbf{Z})}(Y - \mathscr{g}(d^i, \mathbf{U}, \mathbf{Z}))\Big\}. \quad (6)
\end{aligned}$$

We defer the detailed derivations in the supplementary. Heuristically, we can recover the estimates of θ^i and $\theta^{i|j}$ (denoted as $\hat{\theta}^i_w$ and $\hat{\theta}^{i|j}_w$) from $\mathbb{E}\left[\psi^i(W, \vartheta, \varrho) \mid_{\vartheta=\theta, \varrho=\rho}\right] = 0$ and $\mathbb{E}\left[\psi^{i|j}(W, \vartheta, \varrho) \mid_{\vartheta=\theta, \varrho=\rho}\right] = 0$ respectively, which are:

$$\begin{aligned}
\hat{\theta}^i_w &= \frac{1}{N}\left\{\sum_{m=1}^{N} \hat{g}(d^i, \mathbf{u}_m, \mathbf{z}_m) + \sum_{m=1}^{N_i} \frac{(y^i_m - \hat{g}(d^i, \mathbf{u}^i_m, \mathbf{z}^i_m))}{\hat{P}_i(\mathbf{x}^i_m, \mathbf{z}^i_m)}\right\}, \quad (7)\\
\hat{\theta}^{i|j}_w &= \frac{1}{N_j}\left\{\sum_{m=1}^{N_j} \hat{g}(d^i, \mathbf{u}^j_m, \mathbf{z}^j_m) + \sum_{m=1}^{N_i} \frac{\hat{P}_j(\mathbf{x}^i_m, \mathbf{z}^i_m)}{\hat{P}_i(\mathbf{x}^i_m, \mathbf{z}^i_m)}\left[y^i_m - \hat{g}(d^i, \mathbf{u}^i_m, \mathbf{z}^i_m)\right]\right\}. \quad (8)
\end{aligned}$$

We call $\hat{\theta}^i_w$ and $\hat{\theta}^{i|j}_w$ in (7) and (8) the IwC estimates. They are regularization unbiased when the residuals between the observed Y and the estimate of $g(d^i, \cdot, \cdot)$ are used as the regularization term. Besides, they are the consistent estimates (see the Remark 3). Theoretically, we can study the consistency using *error decomposition* (see the supplementary). We also study the consistency with numerical results in the Section "Experiments".

Remark 3. *$\hat{\theta}^i_w$ and $\hat{\theta}^{i|j}_w$ are the consistent estimates of θ^i and $\theta^{i|j}$ if $\hat{P}_i$ and $\hat{g}$ converge to P_i and g sufficiently well in probability (e.g., at rate $N^{-\frac{1}{4}}$) when N and N_j tend to infinity.*

Whenever the estimates of θ^i and $\theta^{i|j}$, i.e., $\hat{\theta}^i$ and $\hat{\theta}^{i|j}$ are available, we can estimate the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATTE) as

$$\begin{aligned}
\mathrm{A\hat{T}E}(i, k) &:= \hat{\theta}^{i,k} = \hat{\theta}^i - \hat{\theta}^k\\
\mathrm{A\hat{T}TE}(i, k|j) &:= \hat{\theta}^{i,k|j} = \hat{\theta}^{i|j} - \hat{\theta}^{k|j}.
\end{aligned} \quad (9)$$

For the IoC formulation, the estimates of ATE and ATTE are denoted as $\hat{\theta}^{i,k}_o$ and $\hat{\theta}^{i,k|j}_o$ which can be computed using (3). For IwC formulation, they are $\hat{\theta}^{i,k}_w$ and $\hat{\theta}^{i,k|j}_w$ computed by (7) and (8) respectively.

Experiments

We now set out experiments to i) estimate the counterfactuals and the treatment effects under different settings and ii) assess the consistency and robustness properties under the IwC formulation.

Our comparisons are made across two aspects: 1) different data properties per (10a) & (10d), and 2) different choices of the map $\hat{g}$ per (2a). For 1), we generate simulated datasets that possess three main properties: a) different levels of causal effect the intervention D causes on the outcome Y, b) different degrees of the nonlinearity of this causal impact, and c) different tail heaviness in the distribution of the feature set $(\mathbf{X}, \mathbf{U}, \mathbf{Z})$. For 2), the various maps under testing include three most commonly used neural network nonlinear models: the Multi-layer Perception Network (MLP), the Convolutional Neural Network (CNN), and the Gated Recurrent Unit (GRU); we also contrast them with widely recognized classic models such as the ordinary least square (OLS) and OLS with LASSO and RIDGE regularization, and the decision-tree based models such as the Random Forest (RF) and one with boosting features, the xgboost (XGB).

For every set of simulated data and a given choice of g, we report estimations of ATE & ATTE per (9), using both our IwC estimations (7) (8) and the IoC estimations (3). All results are out-of-sample and we use 70% of data as the training set and the remaining 30% as the testing set. We use the grid search to find the optimal hyperparameters of the linear models and the tree-based models. For all neural network-based models, we use the Bayesian optimization to find the optimal hyperparameters. The number of hidden layers ranges from 2 to 7 and the number of units for each layer from 50 to 500. The batch size is in integer multiples of 32 and optimized within $[32, 3200]$. We search the learning rate between 0.0001 and 0.1. The experiments are run on two Ubuntu HP Z4 Workstations each with Intel Core i9 10-Core CPU at 3.3GHz, 128G DIMM-2166 ECC RAM, and two sets of NVIDIA Quadro RTX 5000 GPU. The total computation time of Table 1 and Table 3 is 177 hours, including all different sets of α and β, with each set containing 8 models; Figure 1 and Figure 2 cost 163 hours in total.

The Data Generating Process. As the ground truth of the counterfactuals is unavailable, we construct a data generating process (DGP) for our credit-related dataset similar to many causal learning works:

$$Y = f(D)q(\mathbf{U}, \mathbf{Z}) + \xi, \tag{10a}$$

$$\begin{aligned} q(\mathbf{U}, \mathbf{Z}) = \Big\{ &\exp\left(\left|\mathbf{a}_0^T \mathbf{Z}\right|\right) \\ &+ e_1 \log\left(e_2 + k(\mathbf{Z})^2 + |k(\mathbf{U})|^\tau\right)\Big\}^r, \end{aligned} \tag{10b}$$

$$D = \sigma\left(\lambda\left(\mathbf{a}_1^T \mathbf{X} + \left|\mathbf{a}_2^T \mathbf{Z}\right|^\gamma + \frac{b_1\ X_9}{1 + |X_3|} + \frac{b_2\ X_{10}}{1 + |X_6|}\right) + \nu\right), \tag{10c}$$

$$f(D) = \alpha + (1-\alpha) \times \left[\beta D^m + (1-\beta)\exp(D^n)\right], \tag{10d}$$

where $k(\mathbf{Z})$ and $k(\mathbf{U})$ in (10b) are defined as

$$\begin{aligned} k(\mathbf{Z}) &= \log\left(\left|(\mathbf{c}_1^z)^T \mathbf{Z} + \textstyle\sum_{r=2}^4 (\mathbf{c}_r^z)^T \bar{\mathbf{Z}}_r\right|\right) \\ &\quad + (\mathbf{c}_1^z)^T \log|\mathbf{Z}| + \textstyle\sum_{r=2}^4 (\mathbf{c}_r^z)^T \log|\bar{\mathbf{Z}}_r|, \\ k(\mathbf{U}) &= \log\left(\left|(\mathbf{c}_1^u)^T \mathbf{U} + \textstyle\sum_{r=2}^4 (\mathbf{c}_r^u)^T \bar{\mathbf{U}}_r\right|\right) \\ &\quad + (\mathbf{c}_1^u)^T \log|\mathbf{U}| + \textstyle\sum_{r=2}^4 (\mathbf{c}_r^u)^T \log|\bar{\mathbf{U}}_r|. \end{aligned} \tag{11}$$

The function σ maps the features to the intervention variable D such that D takes five treatment levels $\left\{d^1, d^2, d^3, d^4, d^5\right\}$. The confounding feature set is a 20-dimensional random vector $\mathbf{Z} = [Z_1, \cdots, Z_{20}]^T$. Simultaneously, the outcome-specific feature set is a 10-dimensional random vector $\mathbf{U} = [U_1, \cdots, U_{10}]^T$ and the intervention-specific feature set is a 10-dimensional random vector $\mathbf{X} = [X_1, \cdots, X_{10}]^T$. All $(\mathbf{Z}, \mathbf{U}, \mathbf{X})$ are correlated random vectors with the correlation matrix parameterized by $C_{ij} = a + (1-a)\exp(-b|i-j|), a \in [0,1], b \in \mathbb{R}^+$. The parameter values of a, b used to generate the correlation matrix, $(\lambda, \gamma, b_1, b_2)$ in (10c), $(\mathbf{a}_0, \tau, r, e_1, e_2)$ in (10b) , $(\mathbf{a}_1, \mathbf{a}_2)$ in (10c), and $(\{\mathbf{c}_i^z\}_{i=1}^4, \{\mathbf{c}_i^u\}_{i=1}^4)$ in (11) are deferred in the supplementary due to space constraints. The quantity $\bar{\mathbf{Z}}_r$ (or $\bar{\mathbf{U}}_r$) in (11) is a column vector such that each entry is the product of r elements taken from $\mathbf{Z}$ (or $\mathbf{U}$) without repetition. For example, $\bar{\mathbf{Z}}_2 = [Z_1 Z_2, Z_1 Z_3, \cdots, Z_{19} Z_{20}]^T$, $\bar{\mathbf{Z}}_3 = [Z_1 Z_2 Z_3, Z_1 Z_2 Z_4, \cdots, Z_{18} Z_{19} Z_{20}]^T$, etc.

Our DGP is reasonable for the following reasons: 1) In many causal works, the researchers never check if their DGP can appropriately fit the real-world data given in their papers (Hainmueller, Mummolo, and Xu 2019; Hill 2011; Lim 2018). Our DGP fits the highly nonlinear and correlated real-world credit data very well. For example, when we fit the real-world dataset (the one used in our semi-synthetic experiment) to (10c), the relative mean square error (MSE) we obtained is 7.9% which is very small. 2) Our DGP is generic enough. The functions in our DGP, such as the power/exponential/logarithm/inverse, are all out of lengthy testings of each individual's feature separately. 3) It allows us to control the degrees of causality and nonlinearity. First, the parameter $\alpha \in [0,1]$ in (10d) controls the amount of the policy impact. The bigger α is, the smaller the impact. In the case $\alpha = 1$, Y is no longer a function of the intervention. Second, the parameters $\beta \in [0,1]$ and $n, m \in \mathbb{R}^+$ control the degree of nonlinearity for a fixed α. For instance, when $m = 1, n = 2$, the smaller β is, the larger the contribution from the term $\exp(D^n)$, hence the more degree of nonlinearity.

Causality vs. Nonlinearity. Table 1 compares the ATE estimations comprehensively across different choices of $\hat{g}$ and different data properties per (10a) & (10d). Each number in column a) "IoC" and column b) "IwC" is a weighted average of the ATE estimation in relative errors w.r.t. the true ATE, computed from $40,000$ observations and 100 independent experiments for each individual setting:

$$\frac{1}{M}\sum_{m=1}^{M} \sum_{\substack{1 \le i,k \le n \\ i \ne k}} \left[\frac{\left|\theta^{i,k;m}\right|}{\sum_{\substack{1 \le i,k \le n \\ i \ne k}} |\theta^{i,k;m}|} \left| \frac{\hat{\theta}^{i,k;m}}{\theta^{i,k;m}} - 1 \right| \right], \tag{12}$$

where M is the number of experiments conducted, $\theta^{i,k;m}$ is the true value of $\theta^{i,k}$ in the m^{th} experiment, and $\hat{\theta}^{i,k;m}$ is the estimate of $\theta^{i,k}$ in the m^{th} experiment based on IwC (or IoC). The ATTE cases are deferred in the supplementary.

Table 1 reports the results when the α and β in (10d) are fixed at different values. When we fix $\alpha = 5\%$, the amount of impact of the intervention D on the outcome Y is large. We call it the "strong causality" case. When the causal effects are strong, whether the data are light-tail distributed or heavy-tail distributed, mostly linear ($\beta = 95\%$) or mostly nonlinear ($\beta = 5\%$), and whether the choice of $\hat{g}$ is a linear model (OLS, LASSO, RIDGE), a tree-based model (RF, XGB) or a neural-network-based model (GRU, CNN, MLP), the IwC estimates always give superior ATE estimations. Similar observations can be made in the "strong nonlinearity" case in Table 1. In this case, we fix $\beta = 5\%$, meaning that the impact of the intervention D on the outcome Y is very nonlinear, irrespective of whether the amount of the impact is larger or smaller.

Even if one uses misspecified linear models such as the OLS, LASSO, and RIDGE on heavy-tail distributed highly-nonlinear data, there is at least a 30% reduction in the estimation errors. When the causal effects and nonlinearity are both strong in the data and one did use the right nonlinear models, e.g., RF, XGB, GRU, CNN, MLP, the error can be reduced by at least 82.7% for the light-tail data and 75.8% for the heavy-tail data. The significant superiority of IwC vs. IoC w.r.t. the ATE estimation across all our settings suggests the importance of considering the "action-response" relationship.

Consistency of the Estimators. We demonstrate the consistency of the estimators through numerical experiments. We repeat the simulated experiment 100 times, then compute two quantities. The first quantity requires us to compute the values of θ^i and $\hat{\theta}^i_w$ based on 100 experiments and find the corresponding relative errors for each i. We then find the mean of all the relative errors. The second quantity is the average of the standard deviation of the difference between θ^i and $\hat{\theta}^i_w$ of 100 experiments. Indeed, the first quantity is computed as (13a) and the second quantity is computed as (13b) which are stated as follows:

$$\frac{1}{K}\sum_{i=1}^{K}\left|\frac{\sum\limits_{m=1}^{M}\hat{\theta}_w^{i;m}}{\sum\limits_{m=1}^{M}\theta^{i;m}}-1\right| \tag{13a}$$

$$\frac{1}{K}\sum_{i=1}^{K}\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}\left([\hat{\theta}_w^{i;m}-\theta^{i;m}]-\frac{1}{M}\sum_{s=1}^{M}[\hat{\theta}_w^{i;s}-\theta^{i;s}]\right)^2}. \tag{13b}$$

Here K and M in (13a) and (13b) are the number of estimators and the number of experiments respectively.

To show that the estimators are consistent, we compare the mean of all the relative errors versus the number of observational data and summarize the results in Figure 1. We notice that when the number of observational data increases, the computed values in each plot becomes smaller except using linear regressors in computing the mean relative error between θ^i and $\hat{\theta}^i_w$. It implies that using linear regressors would cause biased estimations but not the case of using nonlinear regressors. This matches with our expectations since the generated dataset is nonlinear. The consistency analysis of $\hat{\theta}^{i|j}_w$ is deferred in the supplementary.

Testing on Real-world Data. We now apply our method to a unique real-world dataset kindly provided by JD Digits, one of the largest global technology firms that operates in both the e-commerce business and the lending business.

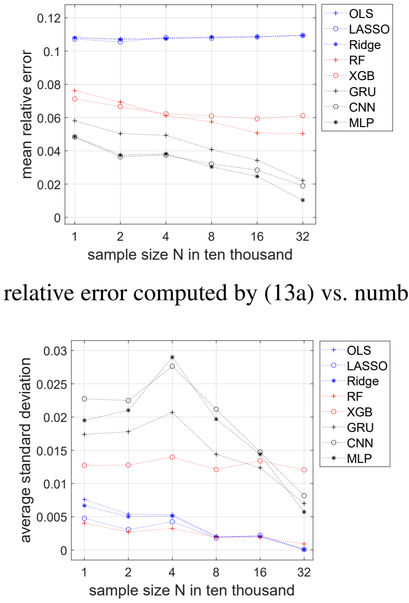
Light-tail experiment

(α, β)	$(95\%, 5\%)$			$(5\%, 5\%)$			$(5\%, 95\%)$		
Method	IoC	IwC	Red.	IoC	IwC	Red.	IoC	IwC	Red.
OLS	45.4	24.5	46.0	44.9	21.5	52.2	42.5	25.0	41.2
LASSO	45.4	17.2	62.0	44.9	21.0	53.2	42.5	18.6	56.3
RIDGE	45.4	24.1	46.8	44.9	21.3	52.5	42.5	24.6	42.0
RF	66.9	10.5	84.3	67.5	8.99	86.7	67.4	10.7	84.1
XGB	88.1	12.0	86.3	87.5	10.1	88.5	88.1	12.7	85.6
GRU	50.8	17.0	66.5	20.5	2.64	87.1	49.5	15.9	67.9
CNN	45.0	15.1	66.5	17.7	3.06	82.7	43.9	14.6	66.8
MLP	47.1	15.0	68.0	16.1	2.35	85.4	45.3	14.3	68.3

Heavy-tail experiment

(α, β)	$(95\%, 5\%)$			$(5\%, 5\%)$			$(5\%, 95\%)$		
Method	IoC	IwC	Red.	IoC	IwC	Red.	IoC	IwC	Red.
OLS	87.6	59.9	31.6	82.1	52.6	35.9	88.7	61.9	30.2
LASSO	87.6	49.1	43.9	79.8	47.2	40.8	88.7	52.0	41.3
RIDGE	87.2	59.5	31.8	81.8	52.3	36.0	88.3	61.4	30.4
RF	67.2	13.8	79.4	68.0	10.3	84.9	68.1	15.3	77.5
XGB	86.9	20.4	76.5	87.3	14.5	83.4	86.9	20.4	76.5
GRU	55.1	20.7	62.4	21.7	5.25	75.8	51.5	23.0	55.3
CNN	48.3	17.8	63.2	20.8	3.80	81.8	48.7	17.9	63.3
MLP	48.9	19.8	59.5	18.5	3.01	83.8	44.4	19.8	55.3

Table 1: Comparison of the ATE estimations per (12) with different *causalities* and *nonlinearities* in simulated data with different tail heaviness (light tail & heavy tail). The values reported are the percentage values. For example, 87.6 in the table actually means 87.6%. "Red." is the error reduction computed by |IwC/IoC-1|. Number of observations $N = 40000$. Number of experiments $M = 100$.



(a) The mean relative error computed by (13a) vs. number of observations.

(b) The average standard deviation computed by (13b) vs. number of observations.

Figure 1: Consistency analysis of $\hat{\theta}^i_w$ in (7) with 100 experiments vs. number of observations N. $\alpha = 0.05$ and $\beta = 0.05$ in (10d); $N \in \{1, 2, 4, 8, 16, 32\} \times 10000$.

Variables	mean	std	min	max	5%	25%	50%	75%	95%
Credit line	4751	2383	100	$5e4$	600	3000	4800	6000	9000
Max pay	507	2695	0	$10e5$	0	0	118	319	2499
Total price	1134	$1.5e4$	0	$6.8e6$	0	0	199	797	4398
Total order	3227	$4.1e4$	0	$2.5e7$	56	347	1116	3269	$1.2e4$
Order days	3.0	4.5	0	87	0	0	2	4	11

Table 2: The mean, standard deviation, minimum, maximum and 5%, 25%, 50%, 75%, 95% quantiles of a few selected variables in the data (records are within 3 months).

The dataset contains observational records of $400,000$ concurrent customers as of writing this paper. The feature dimension of the raw data is 1159, including 1) the borrowers' shopping, purchasing and payment histories; 2) credit lines, interest charges and financing application decisions set by the lender's decision algorithm; 3) outstanding amounts and delinquency records. After autoencoding, the dimension of the feature set is reduced to 40, which will be our $(\mathbf{Z}, \mathbf{U}, \mathbf{X})$. Descriptive statistics of a few representative and important features are stated in Table 2.

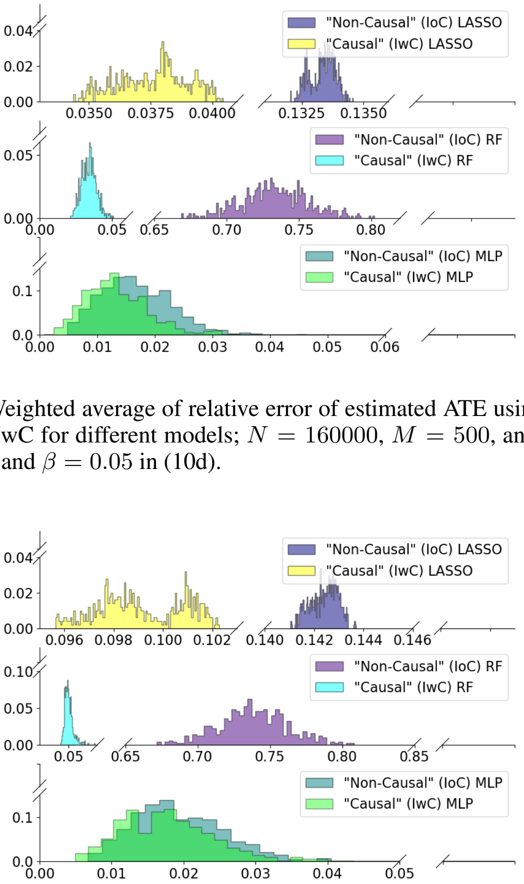
In the field of credit risk analysis, a default outcome is defined w.r.t. a specific credit event. The event in this study is the "three-month delinquent" event. It is the borrowers' payment for any of their outstanding loans within the next three months, no matter the borrowers are late on the payment or not. The intervention variable is the credit line set by the lender's lending algorithm, at the time the customers apply for shopping credit to finance their online purchases.

Different from simulation studies, the true θ^i and $\theta^{i|j}$ are unavailable since the counterfactuals can never be observed in the real application. As such, we adopt the semi-synthetic approach to generate the ground truth of counterfactuals, as commonly used in the literature (e.g., (Hill 2011; Johansson, Shalit, and Sontag 2016; Louizos et al. 2017) and the references therein). The details of the semi-synthetic setup can be found in the supplementary.

The results reported in Figure 2 and Table 3 are strikingly encouraging, given these are from the real-world data. Not only do we see a single-digit percentage estimation error in all settings whenever the IwC estimates are used, we see both significant and robust error reductions compared to the IoC estimates.

Conclusion

This paper presents the first retail credit risk study that estimates the action-response causality from observational data of both a) the e-commerce lender's credit decision records and b) their borrowers' purchase, borrowing, and payment records. Our study shows that the confounding effects between the lender's credit decisions and the borrowers' credit risks, if overlooked, would result in significant biases in risk assessment. Using our IwC formulations, the biases can be reduced to a few percentages. The larger the dataset is, the higher the error reduction. The nature of the current study is about state estimation. Our future study will be towards the decision making problem built on the current framework.



(a) Weighted average of relative error of estimated ATE using IoC and IwC for different models; $N = 160000$, $M = 500$, and $\alpha = 0.05$ and $\beta = 0.05$ in (10d).

(b) Weighted average of relative error of estimated ATTE using IoC and IwC for different models; $N = 160000$, $M = 500$, and $\alpha = 0.05$ and $\beta = 0.05$ in (10d).

Figure 2: Frequency histogram of weighted average of relative error of estimated ATE and ATTE for three models: LASSO, RF and MLP for 500 experiments.

Semi-synthetic experiment

(α, β)	(50%, 5%)			(5%, 5%)			(5%, 50%)		
Method	IoC	IwC	Red.	IoC	IwC	Red.	IoC	IwC	Red.
OLS	14.0	6.40	54.2	13.6	4.81	64.7	13.8	6.27	54.7
LASSO	14.0	6.14	56.1	13.6	4.73	65.3	13.8	6.21	55.1
RIDGE	14.0	6.36	54.5	13.6	4.78	64.9	13.8	6.23	55.0
RF	73.5	3.97	94.6	73.5	2.95	96.0	74.0	3.67	95.0
XGB	88.4	7.53	91.5	88.2	5.86	93.4	88.3	7.12	91.9
GRU	6.38	2.98	53.2	4.07	2.41	40.8	7.32	3.24	55.7
CNN	4.79	3.10	35.3	3.23	2.20	31.9	4.55	3.38	25.7
MLP	3.72	2.77	25.5	3.19	2.04	36.0	3.68	3.37	8.39

Table 3: Comparison of the ATE estimations per (12) under different *nonlinearities* and *causalities* in semi-synthetic experiment. The values reported are the percentage values. For example, 14.0 in the table means 14.0%. "Red." is the error reduction computed by |IwC/IoC-1|. Number of observations $N = 80000$. Number of experiments $M = 100$.

Acknowledgments

Qi WU acknowledges the support from the JD Digits - CityU Joint Laboratory in Financial Technology and Engineering, the Laboratory for AI Powered Financial Technologies, and the GRF support from the Hong Kong Research Grants Council under GRF 14211316, 14206117 and 11219420.

References

Alaa, A.; and Schaar, M. 2018. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, 129–138.

Bennett, A.; and Kallus, N. 2019. Policy evaluation with latent confounders via optimal balance. In *Advances in Neural Information Processing Systems*, 4827–4837.

Berg, T.; Burg, V.; Gombović, A.; and Puri, M. 2020. On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies* 33(7): 2845–2897.

Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W. K.; and Robins, J. 2018. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal* 21(1): C1–C68.

Djebbari, H.; and Smith, J. 2008. Heterogeneous impacts in PROGRESA. *Journal of Econometrics* 145(1-2): 64–80.

Du, Z.; and Zhang, L. 2015. Home-purchase restriction, property tax and housing price in China: A counterfactual analysis. *Journal of Econometrics* 188(2): 558–568.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 1097–1104.

Farrell, M. H. 2015. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* 189(1): 1–23.

Hainmueller, J.; Mummolo, J.; and Xu, Y. 2019. How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis* 27(2): 163–192.

Heckman, J. J.; Ichimura, H.; and Todd, P. 1998. Matching as an econometric evaluation estimator. *The review of economic studies* 65(2): 261–294.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1): 217–240.

Hirano, K.; Imbens, G. W.; and Ridder, G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4): 1161–1189.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029.

Kallus, N. 2018. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, 8895–8906.

Khandani, A. E.; Kim, A. J.; and Lo, A. W. 2010. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance* 34(11): 2767–2787.

Lattimore, F.; Lattimore, T.; and Reid, M. D. 2016. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, 1181–1189.

Li, K. T.; and Bell, D. R. 2017. Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics* 197(1): 65–75.

Li, S.; and Fu, Y. 2017. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, 929–939.

Lim, B. 2018. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*, 7483–7493.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, 6446–6456.

Mackey, L.; Syrgkanis, V.; and Zadik, I. 2018. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, 3375–3383. PMLR.

Malekipirbazari, M.; and Aksakalli, V. 2015. Risk assessment in social lending via random forests. *Expert Systems with Applications* 42(10): 4621–4631.

Neyman, J. 1979. C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A* 1–21.

Oprescu, M.; Syrgkanis, V.; and Wu, Z. S. 2019. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, 4932–4941.

Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, 2503–2513.

Sirignano, J.; Sadhwani, A.; and Giesecke, K. 2016. Deep learning for mortgage risk. *arXiv preprint arXiv:1607.02470* .

Swaminathan, A.; and Joachims, T. 2015a. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, 814–823.

Swaminathan, A.; and Joachims, T. 2015b. The self-normalized estimator for counterfactual learning. In *advances in neural information processing systems*, 3231–3239.

Syrgkanis, V.; Lei, V.; Oprescu, M.; Hei, M.; Battocchi, K.; and Lewis, G. 2019. Machine learning estimation of heterogeneous treatment effects with instruments. In *Advances in Neural Information Processing Systems*, 15167–15176.

Toulis, P.; and Parkes, D. C. 2016. Long-term causal effects via behavioral game theory. In *Advances in Neural Information Processing Systems*, 2604–2612.

Wang, X.; Zhang, R.; Sun, Y.; and Qi, J. 2019. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, 6638–6647.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, 2633–2643.

Yoon, J.; Jordon, J.; and van der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*.

Yuan, B.; Hsia, J.-Y.; Yang, M.-Y.; Zhu, H.; Chang, C.-Y.; Dong, Z.; and Lin, C.-J. 2019. Improving Ad Click Prediction by Considering Non-displayed Events. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 329–338.