Dear author,

Please note that changes made in the online proofing system will be added to the article before publication but are not reflected in this PDF.

We also ask that this file not be used for submitting corrections.

Neurocomputing xxx (xxxx) xxx

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A Pareto-smoothing method for causal inference using generalized Pareto distribution☆

Q1 Fujin Zhu[a,b], Jie Lu[b,*], Adi Lin[b], Guangquan Zhang[b]

[a] School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China
[b] Centre for Artificial Intelligence, School of Computer Science, Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 9 March 2019
Revised 15 July 2019
Accepted 20 September 2019
Available online xxx

Communicated by Dr. Shohei Shimizu

Keywords:
Causality
Causal inference
Machine learning
Treatment effect
Importance sampling

ABSTRACT

Causal inference aims to estimate the treatment effect of an intervention on the target outcome variable and has received great attention across fields ranging from economics and statistics to machine learning. Observational causal inference is challenging because the pre-treatment variables may influence both the treatment and the outcome, resulting in confounding bias. The classic inverse propensity weighting (IPW) estimator is theoretically able to eliminate the confounding bias. However, in observational studies, the propensity scores used in the IPW estimator must be estimated from finite observational data and may be subject to extreme values, leading to the problem of highly variable importance weights, which consequently makes the estimated causal effect unstable or even misleading. In this paper, by reframing the IPW estimator in the importance sampling framework, we propose a Pareto-smoothing method to tackle this problem. The generalized Pareto distribution (GPD) from extreme value theory is used to fit the upper tail of the estimated importance weights and to replace them using the order statistics of the fitted GPD. To validate the performance of the new method, we conducted extensive experiments on simulated and semi-simulated datasets. Compared with two existing methods for importance weight stabilization, i.e., weight truncation and self-normalization, the proposed method generally achieves better performance in settings with a small sample size and high-dimensional covariates. Its application on a real-world heath dataset indicates its utility in estimating causal effects for program evaluation.

1. Introduction

Recent advances in data collection and storage technologies have made more and more observational data available to researchers and decision makers. In the face of such a big data challenge, automatic machine learning algorithms have been used to elicit knowledge from collected data and then make predictions on new data. Most existing machine learning algorithms aim to achieve high predictive accuracy for a target outcome variable [1]. However, in order to make scientific conclusions and rational decisions, we fundamentally need to answer causal questions [2,3], understand the causal relationships between variables or events [4–6], and estimate the possible changes or difference in outcome caused by a particular treatment or policy variable [7,8]. For instance, in biology, scientists conduct randomized experiments to discover and measure the effect of genes on certain genotypes; in healthcare, patients need to know the potential effect on their health to decide whether to take a particular medication; in economics, policy makers debate the possible effect of job training on employees' earning; and in marketing, what ad companies are really interested is the causal effect of an online advertisement on customers' purchasing habits. In the literature, this problem is called *causal inference* or *treatment effect estimation* [7,8].

It has been argued that the ability to learn causality from data is a significant component of human-level intelligence [3,9], and causal inference is a central topic for both scientific discovery [10,11] and decision-making [2]. Causal inference is the problem of estimating the treatment effect of an intervention on a target outcome variable, which is usually the difference between the treatment and control groups in a randomized control trial (RCT) [7]. With growing interest in using data to guide decision making in domains where interventional and counterfactual questions abound, methods for causal inference have attracted considerable research interest [12,13]. RCTs, also known as A/B testing in online learning, ensure the treatment assignment will not be confounded with measured or unmeasured covariates, and thus are the golden standard for estimating treatment effects. However, in many cases

☆ Communicated by
* Corresponding author.
E-mail addresses: Fujin.Zhu@student.uts.edu.au (F. Zhu), Jie.Lu@uts.edu.au (J. Lu), Adi.Lin@student.uts.edu.au (A. Lin), Guangquan.Zhang@uts.edu.au (G. Zhang).

https://doi.org/10.1016/j.neucom.2019.09.095
0925-2312/

they are expensive, unethical, or even impossible. So we can only conduct observational studies using observational data, but identifying the true treatment effect from observational data is challenging because we can only observe the outcome corresponding to the treatment received by an individual, while outcomes under alternative treatments are unobserved. This is called the *fundamental problem of causal inference* [14]. Moreover, the treatment assignment mechanisms, which are usually dependent on the individual's characteristics, are neither known nor random. This makes observational studies inherently more difficult than studies based on RCTs. In clinical medicine, for example, patients receiving different treatments from one another are likely to exhibit different pre-treatment characteristics that may affect outcomes. This crucial problem in estimating causal effect from observational data is called *confounding bias* [14,15].

To minimize the confounding bias in observational causal inference, Rosenbaum and Rubin [14] introduced the propensity score to summarize the information required to control the confounders. The propensity score is the conditional probability of an individual to be assigned to the treatment group, and its estimation helps researchers to better understand the treatment assignment mechanism [16]. Theoretically, we are able to account for the difference between the treatment and control groups by directly modelling the assignment mechanism with propensity scores, thus making the treated and control populations more comparable. In [14], the authors show that if the treatment assignment is ignorable given the observed covariates, the average treatment effect (ATE) can be consistently estimated by adjusting for the propensity scores alone. Given this balancing and de-biasing property, propensity score-based approaches have been widely used for causal inference from observational data [7], complete-case analysis for missing data [17], and survey sampling (Thompson 2012). They have also recently been adopted by the data mining and machine learning communities for de-biasing in recommender systems [18,19], information retrieval systems [20] and learning to rank systems [21].

Despite their popularity and theoretical appeal, a practical problem of propensity-based methods is that the true propensity scores are intrinsically unknown and must be estimated from finite observational data in pure observational studies. Research indicates that misspecification of the propensity score model can result in substantial bias in causal effect estimation. If the estimated propensity scores are close to one or zero for a substantial fraction of the population, the estimated causal effect may be of high variability and difficult to estimate precisely[12]. This is a particular concern in settings with many covariates, or simply when the assignment mechanism is highly skewed. When many of the estimated propensity scores are close to zero, the distribution of their reciprocals the inverse propensity weights are likely to have a heavy right tail, which leads to unstable estimates of treatment effects, sometimes with infinite variance.

To address the problem of variability, two methods for variance control in importance sampling, weight truncation and weight self-normalization, have been used to stabilize the importance weights-based estimators in the causal inference community [7,8]. Researchers from the sampling and weighting community have proposed a growing list of techniques for variance reduction. For a comprehensive understanding, we refer the reader to Owen [22]. In this paper, we reframe causal inference using the IPW estimator in the importance sampling framework and introduce a new smoothing method for importance weight stabilization using the smoothing property of the generalized Pareto distribution (GPD) from the extreme value statistics [23]. Based on the new interpretation of the IPW estimator and the proposed Pareto-smoothing method, we propose two IPW estimators for treatment effect estimation. The proposed Pareto-smoothing method has the following features: Compared with the truncated IPW estimator, our method is less biased and more data efficient in that it has a higher effective sample size. Compared with the self-normalized IPW estimator, the experiment result shows that they both converge to the true value if there is enough data. A special merit of our method is that it is more stable in small sample size cases, which are common in many real-world observational studies.

Our contributions are as follows: (1) We introduce the classic IPW causal estimator from the perspective of importance weighted estimation of expectations using data from a different proposal distribution. To the best of our knowledge, we are the first to formalize such an interpretation of the IPW estimator, which renders the high variability problem of importance weight-based estimators straightforward and easy to understand. (2) Building upon the above importance sampling interpretation of the IPW estimator, we analyse the high variability problem of the IPW estimator with estimated propensity scores and conclude two existing stabilization methods for importance weight stabilization, i.e., weight truncation and self-normalization. (3) We propose a new Pareto-smoothing method for importance weight stabilization using GPDs and two Pareto-smoothed causal estimators based on the proposed method. We also discuss the selection of related parameters in the proposed method. Comprehensive experiments were conducted using both simulated and real data to demonstrate the practical validity of the proposed method.

The remainder of the paper is organized as follows. In Section 2, we introduce notations, formalize the causal inference problem, and discuss the assumptions for identification. In Section 3, we reframe the classic IPW estimator for causal inference in the importance sampling framework, which leads to a straightforward understanding of its high variability problem in finite-sample settings. Within this framework, we briefly review two conventional methods for stabilizing the IPW estimator. In Section 4, we introduce the details of our new Pareto-smoothing method and the two proposed Pareto-smoothed causal estimators. Experiments on simulated data and an application on a real-world health dataset are conducted in Sections 5 and 6. Section 7 concludes the paper and discusses future work.

2. Problem formulation

Consider a population of n individuals, indexed by $i = 1, 2, \ldots n$. Every individual i is characterized by a d-dimensional vector of features (also called pre-treatment covariates or attributes), $X_i \in \mathbb{R}^d$. Elements of these covariates might include age, gender, race, education, etc. In this paper, we use X_i and i interchangeably to represent the ith individual, and X to represent a general individual from the population. Each individual makes a decision to choose an action or is assigned to a treatment T; for example, the treatment T could be whether to take a particular medicine or whether to receive a certain training program. In this paper, we consider binary treatments and denote the treatment for an individual i as T_i, where $T_i = 0$ indicates that individual i received the control treatment and $T_i = 1$ indicates that individual i received the active treatment. Let Y be the outcome variable of interest. For any individual X, following Rubin's potential outcome framework [7], there is a pair of potential outcomes $Y_X(0)$ and $Y_X(1)$, denoting the outcome value of X if he or she had been in the control group or the treatment group respectively. By the principle of consistency, the observed outcome of individual X_i, denoted as $Y_{X_i}^{obs}$ or simply Y_i, is the potential outcome corresponding to the received treatment, i.e., $Y_{X_i}^{obs} = Y_i = Y_{X_i}(T_i) = Y_i(T_i)$

With these notations, the individual treatment effect for the ith individual is defined as the difference of the two potential outcomes $\tau_i = Y_i(1) - Y_i(0)$. The conditional average treatment effect is defined as $\tau(\mathrm{x}) = \mathbb{E}[\tau_i | X_i = \mathrm{x}]$ and the ATE of treatment T on the

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

outcome Y is its expectation for this population,

$$\tau_{ATE} = \mathbb{E}[\tau(X)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \quad (1)$$

Rather than the ATE for the whole population, sometimes we may only be concerned about the ATE for the treated individuals, i.e., the average treatment effect on the treated (ATT) defined as $\tau_{ATT} = \mathbb{E}[\tau(X) \mid T_i = 1]$. While we can analogously define the average treatment effect on the control (ATC), it is seldom of interest in practical applications. We formulate the ATE estimation problem for concreteness. The estimation of ATT is straightforward and is introduced in the Appendix.

Given the observational data $\mathcal{D} = \{(X_i, T_i, Y_i) : i = 1, 2, \ldots, n\}$, where n is the number of observations. Referring to the individuals with $T_i = 1$ as treated individuals and the individuals $T_i = 0$ as control, we also denote the number of treated as $n_1 = \sum_{i=1}^{n} T_i$ and the number of controls as $n_0 = \sum_{i=1}^{n} (1 - T_i)$. For each $i = 1, 2, \ldots, n$, $Y_i(T_i) = Y_i$ is the observed factual outcome and $Y_i(1 - T_i)$ is the counterfactual outcome, i.e., the outcome for individual i had she received the treatment $(1 - T_i)$ instead of T_i. If we have access to both potential outcomes, ATE can be estimated by

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}_X[\mathbb{E}[Y_X(1)] - \mathbb{E}[Y_X(0)]] \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)\end{aligned} \quad (2)$$

The ATE measures the average causal difference of a population if *all* individuals are treated versus *all* are untreated, which is generally different from the conditional difference between the outcomes of the treated group and the control group in the observational data. As a baseline, we denote the empirical conditional difference calculated in Eq. (3) as a naive ATE estimator,

$$\hat{\tau}_{ATE}^{\text{Naive}} = \frac{1}{n_1}\sum_{i=1}^{n} T_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n} (1 - T_i) Y_i \quad (3)$$

Estimating the ATE from observational data is generally impossible because of *the fundamental problem of causal inference* [7]: for each individual, only one of the potential outcomes is observed. As a result, causal inference from observational data is by nature a missing data problem [24]. To ensure the identifiability [25], throughout the paper, we assume *unconfoundedness* (or *conditional exchangeability*) as in Assumption 1.

Assumption 1 (Unconfoundedness, or conditional exchangeability)**.** Conditional on the observed pre-treatment covariates X, the potential outcomes $Y_X(0)$, $Y_X(1)$ are independent of the treatment T, i.e.,$\{Y_X(0), Y_X(1)\} \perp\!\!\!\perp T|X$

This is to say that all confounders that affect both the treatment and outcome are observed. Under this assumption, the backdoor adjustment criterion [25] suggests that we can identify the expected potential outcome $\mathbb{E}[Y_X(t)]$ by the conditional mean outcome via

$$\mathbb{E}[Y_X(t)] = \mathbb{E}[Y|X, T = t]$$

As a result, we can fit two conditional mean outcome models $\mathbb{E}[Y|X_i, T_i = 0]$ and $\mathbb{E}[Y|X_i, T_i = 1]$ from the observational data $\mathcal{D}$, and estimate the ATE in Eq. (2) by calculating the average of the covariate-stratified differences weighted by the probabilities of each stratum. Although feasible in principle, adjusting for all observed covariates to eliminate confounding bias may not be possible, especially when the covariates are continuous. So we need to find a lower-dimensional proxy for them that will suffice for removing the bias associated with imbalance in the pre-treatment covariates.

The propensity score in Definition 1 is such a low-dimensional proxy and plays a key role in many existing propensity score-based causal estimators.

Definition 1 (Propensity score [7])**.** The **propensity score**, $e(X)$, of an individual X is its conditional probability to be assigned to the treatment group, i.e., $e(X) = p(T = 1 \mid X)$.

For any individual, the treatment assignment T is independent of the pre-treatment covariates X conditional on the true propensity score $e(X)$. Moreover, the unconfoundedness assumption implies that $\{Y_X(0), Y_X(1)\} \perp\!\!\!\perp T|e(X)$. In practice, to guarantee enough randomness in the data-generating process so that unobserved counterfactuals can be estimated from the observed data, we also make the *Positivity* assumption.

Assumption 2 (Positivity, or overlap)**.** $0 < e(X_i) < 1$ for any $i = 1, \ldots, n$

This assumption means that the treatment assignment is not deterministic. In words from the literature of observational studies, the observations are generated by a probabilistic assignment mechanism [7].

3. Related work

As we can see from Eq. (1), a key task for treatment effect estimation is to estimate the expected potential outcomes of the population, $\mathbb{E}[Y(0)]$ and $\mathbb{E}[Y(1)]$. In this section, we introduce importance weighted expectation estimators from the importance sampling literature [22]. Within this importance weighting framework, we further introduce the IPW estimator, the truncated and the self-normalized estimators for causal inference.

3.1. Estimating expected potential outcomes

To explain the deduction, let us first consider treatment effect estimation via RCTs and imagine there is a randomized control experiment in which the treatment propensity $p(T_i = 1|X_i)$ is constant for any $i = 1, 2, \ldots, n$. Using the Bayes rule, we can easily derive that the covariate distribution for the treated group, $p_X^{t=1} := p(X|T = 1)$, and the control group, $p_X^{t=0} := p(X|T = 1)$, all equals the population distribution, $p_X := p(X)$. Thus, we can identify both expected potential outcomes via

$$\mathbb{E}[Y(1)] = \mathbb{E}_X \mathbb{E}[Y|X, T = 1] = \mathbb{E}_{p_X^{t=1}} \mathbb{E}[Y|X, T = 1] \quad (4)$$

$$\mathbb{E}[Y(0)] = \mathbb{E}_X \mathbb{E}[Y|X, T = 0] = \mathbb{E}_{p_X^{t=0}} \mathbb{E}[Y|X, T = 0] \quad (5)$$

As a result, ATE can be directly identified from the experimental data via

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \mathbb{E}_{p_X^{t=1}} \mathbb{E}[Y|X, T = 1] - \mathbb{E}_{p_X^{t=0}} \mathbb{E}[Y|X, T = 0]\end{aligned}$$

However, in observational studies, the treatment assignment is generally not random, i.e., $p_X^{t=1} \neq p_X$ and $p_X^{t=0} \neq p_X$. Consequently, we cannot calculate the population expectations $\mathbb{E}[Y(1)]$ and $\mathbb{E}[Y(0)]$ from the observed data directly via Eqs. (4) and (5). Importance sampling is one of the most generally applicable procedures for computing expectations when it is not possible to sample directly from the target distribution. Denote the target distribution as $\pi(x)$ and a proposal distribution $q(x)$. The expectation of any function $h(x)$ with respect to the target distribution $\pi(x)$ can be consistently estimated by the following importance weighting formula [22]

$$\mathbb{E}_\pi[h(x)] = \int h(x)\pi(x)dx = \int h(x)q(x)\frac{\pi(x)}{q(x)}dx$$

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

Denote $w(\mathrm{x}) = \pi(\mathrm{x})/q(\mathrm{x})$ and call $w(\mathrm{x}^s) = \pi(\mathrm{x}^s)/q(\mathrm{x}^s)$ the *importance weight* for the sth sample. If we have S draws $\{\mathrm{x}^1, \mathrm{x}^2, \ldots, \mathrm{x}^S\}$ from $q(x)$, then we can approximate $\mathbb{E}_\pi[h(\mathrm{x})]$ using Monte Carlo by

$$\begin{aligned}\mathbb{E}_\pi[h(\mathrm{x})] &= \int q(\mathrm{x})h(\mathrm{x})\frac{\pi(\mathrm{x})}{q(\mathrm{x})}d\mathrm{x} \\ &= \mathbb{E}_q[w(\mathrm{x})h(\mathrm{x})] \\ &= \frac{1}{S}\sum_{s=1}^{S} w(\mathrm{x}^s)h(\mathrm{x}^s)\end{aligned} \tag{6}$$

In our causal inference setting, the observational data $\mathcal{D} = \{(X_i, T_i, Y_i) : i = 1, 2, \ldots, n\}$ comes from the propensity model $p(T_i = 1|X_i) = e(X_i)$, $p(T_i = 0|X_i) = 1 - e(X_i)$ and the outcome model $Y_i = Y_{X_i}(T_i)$. Knowing that $p(T_i = 1) = \frac{n_1}{n}$, using the above importance weighting formula Eq. (6) and the Bayes rule, we can consistently estimate the expected treated outcome for the population by

$$\begin{aligned}\mathbb{E}[Y(1)] &= \frac{1}{n_1}\sum_{i:T_i=1}\frac{p(X_i)}{p(X_i|T_i=1)}Y_i \\ &= \frac{1}{n_1}\sum_{i:T_i=1}\frac{p(T_i=1)}{p(T_i=1|X_i)}Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}[T_i=1]}{p(T_i=1|X_i)}Y_i\end{aligned} \tag{7}$$

where $\mathbb{1}[T_i = t]$ is the indicator function. Similarly, the expected control outcome for the population can be estimated by

$$\mathbb{E}[Y(0)] = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}(T_i=0)}{p(T_i=0|X_i)}Y_i \tag{8}$$

3.2. IPW estimator

Substituting Eqs. (7) and (8) into the ATE definition in Eq. (2), we get the following ATE estimator

$$\begin{aligned}\hat{\tau}_{ATE} &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}[T_i=1]}{p(T_i=1|X_i)}Y_i - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}[T_i=0]}{p(T_i=0|X_i)}Y_i\end{aligned} \tag{9}$$

Define the importance weight, W_i, for individual i in a general form as the reciprocal of its probability of receiving the observed treatment T_i. Formally,

$$W_i := \frac{1}{p(T_i|X_i)} = \frac{\mathbb{1}(T_i=1)}{e(X_i)} + \frac{\mathbb{1}(T_i=0)}{1-e(X_i)} \tag{10}$$

Then we can rewrite the estimator in Eq. (9) as the following importance weighting estimator

$$\begin{aligned}\hat{\tau}_{ATE}^{\mathrm{IPW}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=1]W_iY_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=0]W_iY_i \\ &= \frac{1}{n}\sum_{i:T_i=1}W_iY_i - \frac{1}{n}\sum_{i:T_i=0}W_iY_i\end{aligned} \tag{11}$$

This is called the IPW estimator [26] and is one of the most commonly used unbiased estimators for treatment effect estimation. In observational studies, the propensity score $e(X_i)$ for each individual is not available and need to be estimated from data by some statistical procedure (for example, Logistic regression). By using the estimated propensity scores $\hat{e}(X_i)$ directly, the finite-sample performance of the IPW estimator $\hat{\tau}_{ATE}^{\mathrm{IPW}}$ could be poor. The reason is that the estimated propensity scores $\hat{e}(X_i)$ occur in the denominator in the definition of importance weight in Eq. (10), and small inaccuracies in $\hat{e}(X_i)$ can induce very high inaccuracies in the estimated ATE, especially when $\hat{e}(X_i)$ is close to zero or one. In this case, the importance weights W_i will be of high variability or even have unbounded variance, thus simple substitute estimators based on them may be unstable and misleading.

To remedy the high variability of the estimated importance weights, we introduce two existing methods for importance weighting estimator stabilization adopted from the importance sampling literature [8]: weight truncation and weight self-normalization.

3.3. Truncated IPW estimator

Weight truncation is a common approach for variance reduction in the importance sampling literature [22,27]. For the purpose of causal effect estimation, the truncated IPW estimator is defined as

$$\begin{aligned}\hat{\tau}_{ATE}^{\mathrm{Trunc}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=1]W_i^{\mathrm{Trunc}}Y_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=0]W_i^{\mathrm{Trunc}}Y_i \\ &= \frac{1}{n}\sum_{i:T_i=1}W_i^{\mathrm{Trunc}}Y_i - \frac{1}{n}\sum_{i:T_i=0}W_i^{\mathrm{Trunc}}Y_i\end{aligned} \tag{12}$$

where the truncated importance weight W_i^{Trunc} is derived by truncating the vanilla importance weight W_i by:

$$W_i^{\mathrm{Trunc}} := \begin{cases} a, & \text{if } W_i < a \\ W_i, & \text{if } a \le W_i \le b \\ b, & \text{if } W_i > b \end{cases} \tag{13}$$

A consequence of weight truncation is the introduction of bias in the truncated importance weights, which in turn causes bias in the importance weight-based estimates. Moreover, the truncation thresholds are usually unknown and choosing them relies on experience or intuition. Crump et al. [28] proposed to keep individuals with estimated propensity score within the range [0.1, 0.9]. As a baseline, we follow this heuristic to truncate the importance weights in Eq. (13) by $a = \frac{10}{9}$ and $b = 10$. We denote the truncated IPW estimator with this truncation thresholds as *TruncCrump*. Recently, Yang and Ding [29] proposed to use a smooth weight function to approximate the existing sample truncation. Their method seems theoretically promising. However it requires us to tune the smooth weight function hyper-parameter and no open source code is available for comparison. In addition, Ju et al. [30] proposed a data-adaptive truncation algorithm which adaptively selects the optimal truncation threshold for the estimated propensity scores, but it is especially designed for target maximum likelihood estimators [31,32]. In this paper, we will compare our proposed estimators with the *TruncCrump* estimator and two other truncated IPW estimators in the experiment sections.

3.4. Self-normalized IPW estimator

We can also apply the control variates technique [22] for variance reduction and divide the importance weights by their empirical mean in each treatment group. Denoting the average importance weight for the treated group as $\overline{W}_t := \frac{1}{n}\sum_{i:T_i=1}W_i$ and the average importance weight for the control group as $\overline{W}_c := \frac{1}{n}\sum_{i:T_i=0}W_i$, the self-normalized importance weight for each individual is then defined as

$$W_i^{\mathrm{Norm}} := \mathbb{1}[T_i=0]\frac{W_i}{\overline{W}_c} + \mathbb{1}[T_i=1]\frac{W_i}{\overline{W}_t}$$

By replacing the importance weights W_i in Eq. (11) by the self-normalized importance weights W_i^{Norm}, we get the following

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

self-normalized IPW estimator

$$\hat{\tau}_{ATE}^{\text{Norm}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}[T_i=1]W_i}{\overline{W}_t}Y_i - \sum_{i=1}^{n}\frac{\mathbb{1}[T_i=0]W_i}{\overline{W}_c}Y_i$$
$$= \frac{1}{n}\sum_{i:T_i=1} W_i^{\text{Norm}}Y_i - \frac{1}{n}\sum_{i:T_i=0} W_i^{\text{Norm}}Y_i \quad (14)$$

In general, the self-normalized IPW estimator $\hat{\tau}_{ATE}^{\text{Norm}}$ has lower variance than the original IPW estimator $\hat{\tau}_{ATE}^{\text{IPW}}$. In the experimental study section, we evaluate the performance of the stabilized IPW estimator, which combines our proposed Pareto-smoothing method with the self-normalization method.

4. Pareto-smoothing method for causal inference

In the previous section, we reframe importance weight-based causal estimators from the perspective of importance sampling estimation of expectations. A common phenomenon in the importance sampling literature is that the importance weighting estimator for expectations are subject to the instability problem in settings with finite samples. To cope with this problem so as to establish stable importance weight-based causal estimators, we also introduced weight truncation and self-normalization, leading to the truncated IPW estimator and the self-normalized estimator. As a complementary of these estimator stabilization methods, in this section, we introduce our Pareto-smoothing method for importance weight stabilization. Based on this method, we further propose two ATE estimators: the Pareto-smoothed IPW estimator and the Pareto-smoothed self-normalized IPW estimator.

Our proposed method builds upon results from the extreme value theory [23]. In extreme value statistics, if an unknown distribution function $F(w)$ lies in the "domain of attraction" of an extreme distribution function, then $F(w)$ has a generalized Pareto upper tail. As a result, we can approximate its upper tail by a GPD if the location μ of the tail can increase as the sample size increases.

Within the framework of importance weight-based causal inference, we estimate the importance weight for each individual and obtain the importance weights $\{W_1, W_2, \ldots, W_n\}$. To remedy the influence of extreme weights, rather than truncating the importance weights in a brute-force way, we fit a GPD over the upper tails of the estimated importance weights and smooth them by the fitted GPD. By this smoothing method, we try to stabilize the importance weights while retaining the information of their relative order.

4.1. GPD fitting

The GPD probability density function for a scalar random variable W with parameter $\theta = (\mu, \sigma, \kappa)$ is defined as

$$f(w) = \begin{cases} 1/\sigma\left(1+\dfrac{\kappa(w-\mu)}{\sigma}\right)^{-1/\kappa-1}, & \kappa \neq 0 \\ 1/\sigma e^{-\dfrac{w-\mu}{\sigma}}, & \kappa = 0 \end{cases} \quad (15)$$

where μ is the location parameter, $\sigma > 0$ is the scale, and κ is the shape of the distribution. In addition, we will also use the cumulative density function $F(w; \mu, \sigma, \kappa)$ defined in Eq. (16) to calculate its expected order statistics as the replacement of large importance weights

$$F(w) = \begin{cases} 1-\left(1+\dfrac{\kappa(w-\mu)}{\sigma}\right)^{-1/\kappa}, & \kappa \neq 0 \\ 1-e^{-\dfrac{w-\mu}{\sigma}}, & \kappa = 0 \end{cases} \quad (16)$$

In this section, we describe the procedure for fitting a GPD over the upper tail of the estimated importance weights $\{W_1, W_2, \ldots, W_n\}$. This includes heuristics for choosing the location parameter μ, estimating the positive scale parameter σ and the shape parameter κ. In general, we only consider fitting the parameters with $\kappa \neq 0$.

4.1.1. Selecting μ

The location parameter μ of a GPD $F(w; \mu, \sigma, \kappa)$ determines the cut-point of the ordered importance weights and thus how many importance weights will be smoothed. In this section, we refer to existing literature and propose to choose it heuristically.

In order to obtain asymptotic consistency, Pickands [33] proposed that the lower bound parameter μ should be chosen so that the sample size M of to-be-smoothed weights in the tail increases to infinity while M/n goes to zero. In addition, by extensive empirical comparisons, Vehtari et al. [34] recommended to choose μ so that the sample size M satisfies

$$M = \min(\lfloor 0.2n \rfloor, \lfloor 3\sqrt{n} \rfloor) \quad (17)$$

This is a reasonable heuristic for deciding the location parameter and the empirical study in [35] shows that the majority of results are not sensitive to the choice of M. In this paper, following this routine, we first sort $W_1, W_2, \ldots, W_n$ in an ascending order and obtain the order statistics of these importance weights, $W_{[1]}, W_{[2]}, \ldots, W_{[n]}$ where $W_{[1]} \leq W_{[2]} \leq \cdots \leq W_{[n]}$. Then the location parameter μ is chosen by

$$\hat{\mu} = W_{[n-M]} \quad (18)$$

where M is derived according to Eq. (17).

4.1.2. Estimating k and σ

Having selected the location μ, we now estimate the scale σ and shape k of the GPD over the upper tail $\{W_{[n-M+1]}, W_{[n-M+2]}, \ldots, W_{[n]}\}$. In statistics, for a GPD $F(w; \mu, \sigma, \kappa)$ over $\{W_{[n-M+1]}, W_{[n-M+2]}, \ldots, W_{[n]}\}$, define

$$O_m = W_{[n-M+m]} - \mu, \quad m = 1, 2, \ldots, M$$

then $\{O_1, O_2, \ldots, O_M\}$ follow the GPD $F(w; 0, \sigma, \kappa)$.

There are many methods to estimate k and σ using the M residuals $\{O_1, O_2, \ldots, O_M\}$ in the literature [36]. Among these methods, Zhang and Stephens [37] reparametrized the GPD $F(w; 0, \sigma, \kappa)$ by two parameters (ρ, κ), where $\rho = \kappa/\sigma$. With this reparameterization, we can easily derive the log-likelihood for the samples $\{O_1, O_2, \ldots, O_M\}$ as

$$\ell(\rho, \kappa) = M\log\frac{\rho}{\kappa} - \frac{\kappa+1}{\kappa}\sum_{i=1}^{M}\log(1+\rho O_i) \quad (19)$$

Set the gradient over κ to 0,

$$\nabla_\kappa \ell = -\frac{M}{\kappa} + \frac{\sum_{i=1}^{M}\log(1+\rho O_i)}{\kappa^2} = 0$$

We get

$$\kappa = \frac{1}{M}\sum_{i=1}^{M}\log(1+\rho O_i) \quad (20)$$

Substituting Eq. (20) into Eq. (19), we get the following *profile log-likelihood* function for ρ

$$\ell(\rho) = M\log\frac{\rho}{\kappa} - M(\kappa+1) \quad (21)$$

where κ is a function of ρ as indicated in Eq. (20). Thus, the key is to get an estimate of ρ. Zhang and Stephens [37] proposed to estimate it using the Bayes-flavoured estimation method as

$$\hat{\rho} = \int \rho \cdot \pi(\rho) L(\rho) d\rho \Big/ \int \pi(\rho) L(\rho) d\rho \quad (22)$$

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

where $L(\rho) = e^{\ell(\rho)}$ is the *profile likelihood function* and the prior $\pi(\rho)$ is specified in a way such that the estimates always exist and can be expressed as explicit functions of the observations. Its derivation is sophisticated and out of the scope of this paper. For more details, we refer the readers to Zhang and Stephens [37].

The estimate has a small bias, is highly efficient, and is simple and fast to compute. With an esitmation $\hat{\rho}$ from Eq. (22), the final estimates for κ and σ are given by

$$\hat{\kappa} = \frac{1}{M}\sum_{i=1}^{M}\log\left(1+\hat{\rho}O_i\right), \qquad \hat{\sigma} = \frac{\hat{\kappa}}{\hat{\rho}} \tag{23}$$

4.2. Weight smoothing

The original importance weights, $\{W_i, i = 1, \ldots, n\}$, are smoothed by replacing the M largest weights with the expected values of the order statistics of the fitted GPD $F(w; \hat{\mu}, \hat{\sigma}, \hat{\kappa})$, i.e.,

$$W_{[n-M+m]} = F^{-1}\left(\frac{m-1/2}{M}\right), \quad m = 1, \ldots, M \tag{24}$$

where $F^{-1}(\cdot)$ is the inverse cumulative distribution of the fitted $F(w; \hat{\mu}, \hat{\sigma}, \hat{\kappa})$. Denote the resulting Pareto-smoothed importance weight for X_i as W_i^{PS}, the above weight replacement procedure is equivalent with

$$W_i^{PS} := \begin{cases} W_i & \text{if } W_i \le \hat{\mu} \\ F^{-1}\left(\dfrac{m_i - n + M - 0.5}{M}\right), & \text{otherwise} \end{cases} \tag{25}$$

where m_i is the order number of W_i in the ascendingly sorted importance weights used in the previous section.

By this procedure, we obtain the Pareto-smoothed importance weights $\{W_1^{\mathrm{PS}}, W_2^{\mathrm{PS}}, \ldots, W_n^{\mathrm{PS}}\}$, which are the basis of the Pareto-smoothed IPW estimator ($\hat{\tau}_{ATE}^{\mathrm{PS}}$) and the Pareto-smoothed self-normalized IPW estimator ($\hat{\tau}_{ATE}^{\mathrm{PSNorm}}$) introduced in the following section.

4.3. Estimators

Given a set of n observations $\mathcal{D} = \{(X_i, T_i, Y_i), \ldots, (X_n, T_n, Y_n)\}$, we fit the importance weights by Logistic regression, obtain the Pareto-smoothed importance weights $\{W_1^{\mathrm{PS}}, W_2^{\mathrm{PS}}, \ldots, W_n^{\mathrm{PS}}\}$ using the above procedures, and estimate the ATE by

$$\begin{aligned}\hat{\tau}_{ATE}^{\mathrm{PS}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=1]W_i^{\mathrm{PS}}Y_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[T_i=0]W_i^{\mathrm{PS}}Y_i \\ &= \frac{1}{n}\sum_{i:T_i=1}W_i^{\mathrm{PS}}Y_i - \frac{1}{n}\sum_{i:T_i=0}W_i^{\mathrm{PS}}Y_i\end{aligned} \tag{26}$$

The process of ATE estimation by our proposed estimator is summarized in Algorithm 1. The corresponding process for ATT estimation is described in the Appendix. In addition, we can also make use of the self-normalization trick after weight smoothing and estimate the ATE by

$$\hat{\tau}_{ATE}^{\mathrm{PSNorm}} = \frac{1}{n}\left(\sum_{i:T_i=1}\frac{W_i^{\mathrm{PS}}}{\overline{W}_t^{\mathrm{PS}}}Y_i - \sum_{i:T_i=0}\frac{W_i^{\mathrm{PS}}}{\overline{W}_c^{\mathrm{PS}}}Y_i\right) \tag{27}$$

where $\overline{W}_t^{PS} = \frac{1}{n}\sum_{i:T_i=1}W_i^{PS}$ and $\overline{W}_c^{PS} = \frac{1}{n}\sum_{i:T_i=0}W_i^{PS}$. This Pareto-smoothed self-normalized IPW estimator proceeds as Algorithm 1 by estimating the ATE using Eq. (27) in the last step.

In general, the Pareto-smoothed IPW Estimator $\hat{\tau}_{ATE}^{\mathrm{PS}}$ stabilizes the IPW estimator with a novel weight smoothing trick. The self-normalized IPW estimator tries to stabilize the estimate by standardizing the importance weights by the average weight in each group. The Pareto-smoothed self-normalized IPW estimator ($\hat{\tau}_{ATE}^{\mathrm{PSNorm}}$) takes advantage of the self-normalization trick used by $\hat{\tau}_{ATE}^{\mathrm{Norm}}$ and further stabilizes $\hat{\tau}_{ATE}^{\mathrm{PS}}$ by standardizing the smoothed importance weights.

Algorithm 1 Pareto-smoothed IPW ATE Estimator.

Input: Observation data $\mathcal{D} = \{(X_i, T_i, Y_i), \ldots, (X_n, T_n, Y_n)\}$
Output: The estimated $\hat{\tau}_{ATE}$

1: Fit the Logistic regression propensity model $e(X) = p(T = 1|X)$ from $\mathcal{D}$;
2: Calculate the importance weights for each individual $\{W_i, i = 1, \ldots, n\}$ via Eq. (10);
3: Sort the importance weights $\{W_i, i = 1, \ldots, n\}$ ascendingly to obtain the sorted importance weights $\{W_{[1]}, W_{[2]}, \ldots, W_{[n]}\}$
4: Choose the location parameter $\hat{\mu}$ by Eq. (18))
5: Estimate the parameters σ and k by Eq. (23))
6: Smooth the importance weights $\{W_1, W_2, \ldots, W_n\}$ by Eq. (25)) to obtain the Pareto-smoothed importance weights $\{W_1^{\mathrm{PS}}, W_2^{\mathrm{PS}}, \ldots, W_n^{\mathrm{PS}}\}$
7: Estimate the ATE $\hat{\tau}_{ATE}$ via Eq. (26)

4.4. Asymptotic analysis

To analyse the asymptotic property of the propoed estimators using results from existing literature, define the weight function

$$\lambda(X_i) := \frac{\mathbb{1}[T_i=1]n_1}{n}\frac{W_i^{\mathrm{PS}}}{\overline{W}_t^{\mathrm{PS}}} + \frac{\mathbb{1}[T_i=0]n_0}{n}\frac{W_i^{\mathrm{PS}}}{\overline{W}_c^{\mathrm{PS}}} \tag{28}$$

We can conclude according to Eq. (25) that $\lambda_i = \lambda(X_i)$ is a function of the covariates X_i parameterized by the propensity model parameters and the fitted GPD parameters. With this notation, we can rewrite $\hat{\tau}_{ATE}^{\mathrm{PSNorm}}$ in Eq. (27) as a standard weighting estimator

$$\hat{\tau}_{ATE}^{\mathrm{PSNorm}} = \frac{1}{n_1}\sum_{i:T_i=1}\lambda_i Y_i - \frac{1}{n_0}\sum_{i:T_i=0}\lambda_i Y_i \tag{29}$$

where the weights λ_i satisfy the following two summation restrictions:

$$\frac{1}{n_1}\sum_{i:T_i=1}\lambda_i = \frac{1}{n}\sum_{i:T_i=1}\frac{W_i^{\mathrm{PS}}}{\overline{W}_t^{\mathrm{PS}}} = 1$$

and

$$\frac{1}{n_0}\sum_{i:T_i=0}\lambda_i = \frac{1}{n}\sum_{i:T_i=0}\frac{W_i^{\mathrm{PS}}}{\overline{W}_c^{\mathrm{PS}}} = 1$$

Define the two conditional variance functions $\sigma_0^2(\mathrm{x}) := \mathbb{V}(Y(0)|X = \mathrm{x})$ and $\sigma_1^2(\mathrm{x}) := \mathbb{V}(Y(1)|X = \mathrm{x})$. According to the results in [28,29] and [7, Chap. 19], if the weighting function $\lambda(X_i)$ is continuous and differentiable, the estimator $\hat{\tau}_{ATE}^{\mathrm{PSNorm}}$ is asymptotic linear and its asymptotic variance can be approximated by

$$\mathbb{V}(\hat{\tau}_{ATE}^{\mathrm{PSNorm}}) \quad = \frac{1}{n_1^2}\sum_{i:T_i=1}\lambda_i^2\cdot\sigma_1^2(X_i) + \frac{1}{n_0^2}\sum_{i:T_i=0}\lambda_i^2\cdot\sigma_0^2(X_i)$$

However, it is easy to verify that the weighting function $\lambda_i = \lambda(X_i)$ in Eq. (28) is not smooth nor differentiable. In this case, we cannot guarantee the consistency of the proposed estimator $\hat{\tau}_{ATE}^{\mathrm{PSNorm}}$. Moreover, the inference of its asymptotic variance is an open problem in the causal inference literature and existing methods are unable to conduct inference to the population [28,29]. Analogously, the Pareto-smoothed IPW estimator $\hat{\tau}_{ATE}^{\mathrm{PS}}$ is also inconsistent. To quantify the estimation uncertainty of the causal estimators, in the simulation and experiment sections, we replicate the experiments multiple times and report the empirical standard error of each estimator.

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

Table 1
Abbrivation (Abbr.) and description of ATE estimators.

Estimator	Abbr.	Description
$\hat{\tau}_{ATE}^{\text{Naive}}$	Naive	Naive estimator for ATE as in Eq. (3)
$\hat{\tau}_{ATE}^{\text{IPW}}$	IPW	IPW estimator for ATE as in Eq. (11)
$\hat{\tau}_{ATE}^{\text{Trunc}}$	Trunc	Truncated IPW estimator for ATE as in Eq. (12) with the truncation thresholds $a = 1$ and b specified by Eq. (17)
	TruncNorm	Truncated IPW estimator for ATE by normalizing the truncated importance weights used in the *Trunc* estimator
	TruncCrump	Truncated IPW estimator for ATE with weight truncation thresholds $a = \frac{10}{9}$ and $b = 10$ in (13)
$\hat{\tau}_{ATE}^{\text{Norm}}$	Norm	IPW estimator for ATE with weight self-normalization for ATE as in Eq. (14)
$\hat{\tau}_{ATE}^{\text{PS}}$ (Ours)	PS	Pareto-smoothed IPW estimator for ATE as in Eq. (26)
$\hat{\tau}_{ATE}^{\text{PSNorm}}$ (Ours)	PSNorm	Pareto-smoothed self-normalized IPW estimator for ATE as in Eq. (27)

We summarize this section by comparing the proposed Pareto-smoothing method with the weight truncation method for causal estimator stabilization. Both methods are biased. while the weight truncation method truncate the extreme weights by fixed values, our proposed method tries to smooth them and keep their relative order. As a result, the proposed Pareto-smoothed estimators are expected to be less biased than truncated estimators. This is validated by the empirical results in the simulation experiments.

5. Simulation studies

Since the ground truth counterfactual outcomes are not available in real-world observational datasets, evaluating causal inference algorithms is not straightforward. In this section, we validate our proposed method using simulated and semi-simulated data, where the ground truth is available to us such that we can evaluate the performance of different methods. Descriptions of all ATE estimators used in the paper are listed in Table 1. Specifically, according to the criterion used for choosing the truncation thresholds, we specify three variants of the truncated IPW estimator. The first truncation estimator, *Trunc*, uses the same criterion, Eq. (17), to specify the truncation thresholds as our Pareto-smoothed estimators. The second truncation estimator, *TruncCrump* uses the truncation threshold in [28] (discussed in Section 3.3). In addition, we also use the self-normalization trick used in the proposed Pareto-smoothed self-normalized IPW estimator to the *Trunc* estimator and denote the resulting estimator the *TruncNorm* estimator. In all the experiments, we follow most of the literature on propensity score estimation and use Logistic regression to fit the propensity scores. The R code for implementing all experiments in this paper is available at https://github.com/dachylong/paretoSmoothing. In all simulations, the underlying potential outcomes $Y_i(0)$ and $Y_i(1)$ for each individual are known, so we can calculate the true sample ATE empirically by $\tau_{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$. For an estimator $\hat{\tau}_{ATE}$, its estimation bias is calculated by

$$Bias_{\text{ATE}} = |\hat{\tau}_{ATE} - \tau_{ATE}| = \left|\hat{\tau}_{ATE} - \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\right|$$

5.1. Simulated data

We simulated data in the context of both low dimensional and relatively high dimensional covariates.

5.1.1. Low dimensional covariates

There are two pre-treatment covariates in the first simulation: one binary $X_{i1}|T_i = 0 \sim$ Bernoulli (0.4),$X_{i1}|T_i = 1 \sim$ Bernoulli (0.5) and one continuous $X_{i2}|T_i = 0 \sim \mathcal{N}(-1.0, 1)$, $X_{i2}|T_i = 1 \sim \mathcal{N}(1.0, 1)$. We simulated data with sample size $n = 100, 200, 300, 500, 1000, 1500, 2000$. For each sample size, we assigned exactly half of the subjects to the treatment group $T = 1$, and the other half to the control group $T = 0$. The potential outcomes for each subject i are adapted from [38] with $Y_i(0) = 0.85X_{i2} + 0.05X_{i2}^2 + 2$ and $Y_i(1) = 0.25X_{i1} + (1 + \exp(1 - 0.85X_{i2}))^{-1}$. With this data generating process, the true ATE is approximately -1.52. For each sample size n, we replicated the experiment 1000 times. The result on ATE estimation biases and standard errors is listed in Table 2. To clearly compare the estimation performance, we also illustrate the estimation bias in terms of the sample size in Fig. 1.

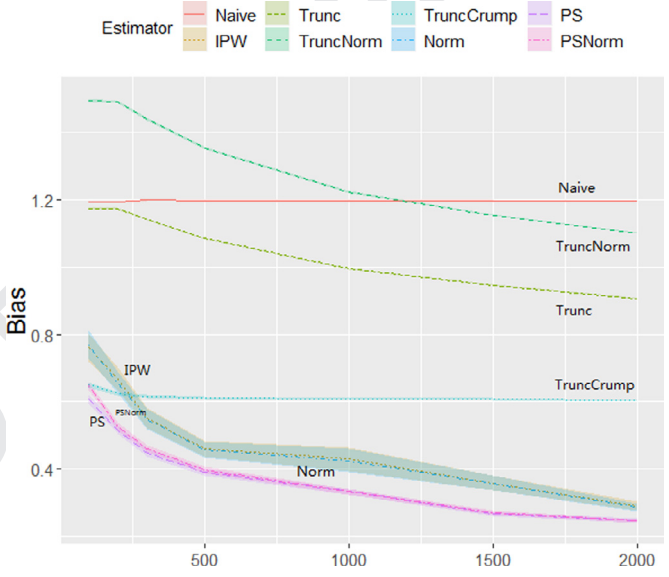

Fig. 1. ATE estimation bias and standard error in terms of sample size n over 1000 replicates for the simulated low-dimensional covariate data.

As we can see from the result, as the sample size increases, all importance weight-based estimators obtain better estimates. In general, estimators based on our proposed Pareto-smoothing method, i.e., the *PS* estimator and the *PSNorm* estimator, achieve the best performance in all sample sizes. As two unbiased estimators, the *IPW* estimator and the *Norm* estimator achieve similar performance. Among the three weight stabilization methods, weight truncation, self-normalization and Pareto-smoothing, our proposed Pareto-smoothing method is the least biased and is more stable than the self-normalization method. By further comparing *TruncNorm* and *Trunc* as well as *PSNorm* and *PS*, we find that self-normalization is likely to worsen the estimation when the sample size is small and one weight stabilization strategy has already been used, either truncation or Pareto-smoothing.

5.1.2. High dimensional covariates

With finite data, the estimated importance weights are more likely to be highly variable in settings with high dimensional

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

Table 2
Comparison of ATE estimation bias and standard error (SE) averaged over 1000 replicates on the simulated low-dimensional covariate data for different estimators list in Table 1.

n	Naive	IPW	Trunc	TruncNorm	TruncCrump	Norm	PS	PSNorm
100	1.194 ± 0.004	0.761 ± 0.042	1.173 ± 0.005	1.494 ± 0.006	0.651 ± 0.008	0.769 ± 0.041	**0.608 ± 0.013**	0.647 ± 0.013
200	1.194 ± 0.003	0.668 ± 0.033	1.173 ± 0.003	1.491 ± 0.004	0.624 ± 0.006	0.656 ± 0.033	**0.515 ± 0.012**	0.526 ± 0.012
300	1.198 ± 0.002	0.551 ± 0.030	1.142 ± 0.003	1.439 ± 0.004	0.616 ± 0.005	0.548 ± 0.029	**0.449 ± 0.011**	0.461 ± 0.010
500	1.197 ± 0.002	0.460 ± 0.023	1.085 ± 0.002	1.354 ± 0.003	0.612 ± 0.004	0.458 ± 0.023	**0.390 ± 0.010**	0.396 ± 0.010
1000	1.197 ± 0.001	0.430 ± 0.035	0.995 ± 0.002	1.223 ± 0.002	0.609 ± 0.002	0.425 ± 0.035	0.334 ± 0.009	**0.333 ± 0.008**
1500	1.198 ± 0.002	0.358 ± 0.022	0.946 ± 0.001	1.154 ± 0.002	0.608 ± 0.002	0.358 ± 0.022	**0.268 ± 0.006**	0.271 ± 0.006
2000	1.196 ± 0.001	0.291 ± 0.012	0.907 ± 0.001	1.100 ± 0.002	0.604 ± 0.002	0.287 ± 0.012	$0.247 \pm .006$	**0.246 ± 0.006**

Table 3
Comparison of ATE estimation bias and standard error (SE) averaged over 1000 replicates on the simulated high-dimensional covariate data for different estimators list in Table 1.

n	Naive	IPW	Trunc	TruncNorm	TruncCrump	Norm	PS	PSNorm
500	1.876 ± 0.014	1.657 ± 0.106	3.869 ± 0.015	2.100 ± 0.012	1.535 ± 0.020	1.774 ± 0.107	**1.300 ± 0.035**	1.388 ± 0.038
1000	1.892 ± 0.010	1.172 ± 0.047	3.353 ± 0.012	1.908 ± 0.010	1.478 ± 0.015	1.270 ± 0.048	**0.990 ± 0.027**	1.064 ± 0.029
1500	1.895 ± 0.008	0.968 ± 0.036	3.064 ± 0.010	1.781 ± 0.008	1.435 ± 0.012	1.057 ± 0.038	**0.839 ± 0.026**	0.907 ± 0.027
2000	1.898 ± 0.007	0.887 ± 0.034	2.887 ± 0.008	1.705 ± 0.007	1.439 ± 0.010	0.965 ± 0.035	**0.737 ± 0.018**	0.799 ± 0.020
2500	1.910 ± 0.006	0.784 ± 0.026	2.765 ± 0.007	1.653 ± 0.006	1.448 ± 0.009	0.860 ± 0.027	**0.697 ± 0.018**	0.758 ± 0.019
3000	1.896 ± 0.006	0.764 ± 0.029	2.641 ± 0.007	1.589 ± 0.006	1.431 ± 0.008	0.829 ± 0.030	**0.660 ± 0.017**	0.713 ± 0.018

covariates. To investigate the performance of the proposed Pareto-smoothing method in this setting, we adapt the simulation in [38] and generate data by assigning half of the samples to the treatment group $T = 1$, and the other half to the control group $T = 0$. In this simulation, there are 10 confounders, 5 binary and 5 continuous. The values of the binary confounders are generated by $X_i|T = 0 \sim \text{Bernoulli}(0.4), X_i|T = 1 \sim \text{Bernoulli}(0.45), i \in \{1, 2, 3, 4, 5\}$ and that of the continuous confounders was generated by $X_i|T = 0 \sim \mathcal{N}(-1, 3^2), X_i|T = 1 \sim \mathcal{N}(1.25, 3^2), i \in \{6, 7, 8, 9, 10\}$. The potential outcomes were generated so that they exhibit non-linear trends in the estimated propensity scores. For each individual i, the two underlying potential outcomes are generated by $Y_i(0) = 5 + 0.2(X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5}) + (1 + \exp(1 - 8X_{i5}))^{-1} + X_{i7} + X_{i8} + X_{i9} + X_{i10}$ and $Y_i(1) = -5 + 0.2(X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5}) - 0.5(X_{i6} + X_{i7} + X_{i8} + X_{i9} + X_{i10})$

We simulate data with sample size $n = 500, 1000, 1500, 2000, 2500, 3000$. The treatment propensity scores are unknown and are estimated via simple Logistic regression, which is clearly misspecified. Estimation biases of different estimators are listed in Table 3. The estimation biases and corresponding standard errors in terms of the sample size are illustrated in Fig. 2. The result is similar with that in the low-dimensional covariate setting, the proposed *PS* estimator obtains the lowest estimation biases in all sample sizes. The proposed *PSNorm* estimator performs slightly worse than *IPW* but better than the other estimators. Results for this high-dimensional covariate setting further validate the superiority of our proposed Pareto-smoothing method.

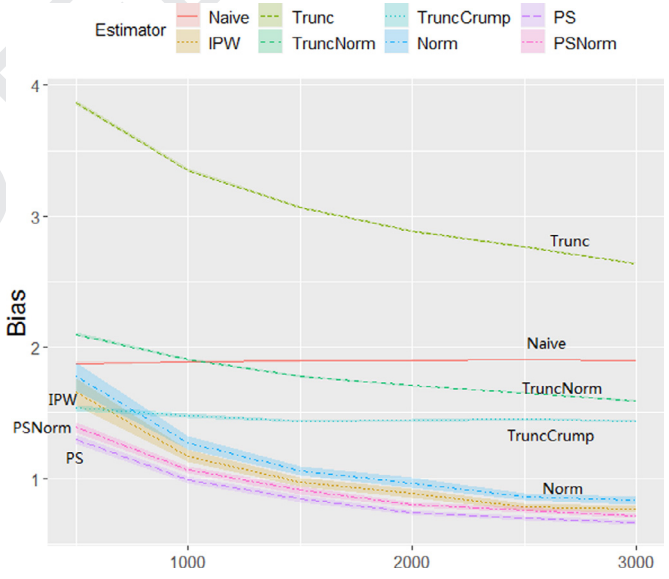

Fig. 2. ATE estimation bias and standard error in terms of sample size n over 1000 replicates for the simulated high-dimensional covariate data.

5.2. Semi-simulated data: IHDP

In this section, we evaluate the performance of our algorithm through the semi-simulated dataset based on the Infant Health and Development Program (IHDP) which was introduced in [39] and used as a benchmark dataset in the causal inference literature [13,40,41]. The IHDP is a real randomized experiment to enhance the cognitive and health status of low birth weight, premature infants through paediatric follow-ups and parent support groups. The observed covariates and treatments in the semi-simulated data are from the IHDP program, while all outcomes (response surfaces) are simulated so that the true treatment effects are known. In total, the IHDP dataset consists of 747 individuals (139 treated, 608 control), and 25 covariates measuring the properties of children and their mothers. The binary treatment T indicates whether the child was assigned into a program where both intensive high-quality childcare and home visits from a trained provider are provided. Examples of covariates include the sex and birth weights of the child, and the age and education attainment level of the mother.

We conduct experiments on all three response simulation settings proposed in [39]. In setting A, the response surfaces are linear and parallel across the two treatment groups and there is no treatment effect heterogeneity. The response surfaces for settings B and C are nonlinear and not parallel across treatment conditions. The outcomes are simulated so that the underlying treatment effects are 4.0. For more details of the three simulation settings, refer to Hill [39]. We simulated the outcomes using the NPCI package[1] and ran the experiment 1000 times. The boxplot of the ATE

[1] https://github.com/vdorie/npci .

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

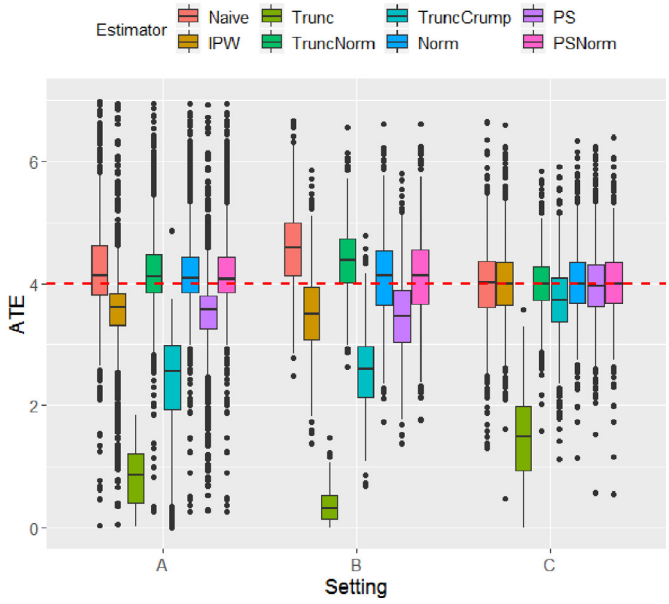

Fig. 3. Box plot of ATEs estimated by different estimators for the IHDP data in different settings. The underlying true ATE is 4 (the red dashed line) in all settings.

Table 4
Results for the IHDP dataset. A, B and C stand for three outcome simulation settings. Estimation biases and standard errors (SE) are computed by replicating the experiment 1000 times.

	A		B		C	
	Bias	SE	Bias	SE	Bias	SE
Naive	0.949	(0.048)	0.711	(0.016)	0.491	(0.014)
IPW	0.744	(0.029)	0.652	(0.015)	0.461	(0.014)
Trunc	6.854	(0.178)	4.989	(0.029)	2.744	(0.030)
TruncNorm	0.733	(0.037)	0.542	(0.013)	0.371	(0.010)
TruncCrump	2.104	(0.059)	1.447	(0.019)	0.519	(0.015)
Norm	0.696	(0.036)	0.539	(0.013)	0.432	(0.013)
PS	0.779	(0.030)	0.678	(0.015)	0.456	(0.013)
PSNorm	0.697	(0.036)	0.538	(0.013)	0.428	(0.012)

estimates of different estimators in three settings is illustrated in Fig. 3. Results of estimation biases are listed in Table 4.

As we can see from Fig. 3, the *Norm* and *PSNorm* estimators perform similarly in all three settings, with estimated ATEs around the true ATE. The estimators *IPW*, *Trunc*, *TruncCrump* and *PS* tend to under-estimate the ATE in setting A and B. Furthermore, the result in Table 4 indicates that the *Norm* estimator and the *PSNorm* estimator achieve the lowest bias in settings A and B respectively. While in setting C, the *TruncNorm* estimator performs the best. In addition, we find that weight truncation or Pareto-smoothing alone deteriorates the ATE estimation performance in settings A and B. The reason may be that since the treatment assignments in the IHDP data are random, the negative influence of weight truncation and Pareto-smoothing proposed for handling extreme importance weights surpasses the benefit they bring for this balanced dataset. Fortunately, by combining them with weight self-normalization, the resulting *TruncNorm* and *PSNorm* estimators acheive better estimation than the naive *IPW* estimator.

6. Application to the NHEFS data

The National Health and Nutrition Examination Survey (NHANES) is a population survey designed to assess the heath and nutritional status of adults and children in the United States. It was jointly initiated by the National Center for Health Statistics (NCHS), which is part of the Centers for Disease Control and Prevention, and the National Institute on Aging in collaboration with other agencies of the US Public Health Service. The datasets, with a detailed description and documentation, are publicly available online.[2] Under the NHANES, the NHANES I Epidemiologic Follow-up Study (NHEFS) was designed to investigate the relationships between clinical, nutritional and behavioural factors assessed in NHANES I.

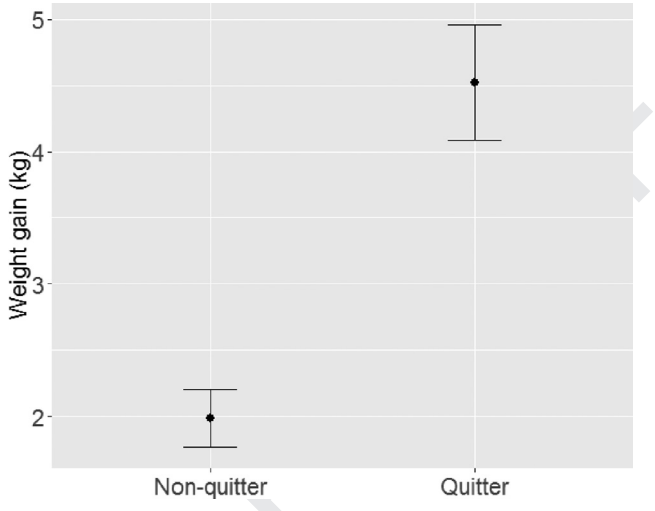

Fig. 4. The mean and standard deviation of weight gains for smoking non-quitters and quitters in the NHEFS data.

We use the subset of the NHEFS dataset used in [8] to estimate the ATE of smoking cessation on weight gain. There are 1746 cigarette smokers in the original data with a baseline visit in the year of 1971–75. After removing missing and censored records, there are 1566 individuals left, aged 25–74 years old and with a follow-up visit in 1982. Individuals who reported having quit smoking before the follow-up visit are classified as treated $T = 1$, and as untreated $T = 0$ otherwise. The outcome variable – weight gain Y – of each individual is the body weight at the follow-up visit minus the body weight at the baseline visit, measured in *kg*. Examples of pre-treatment covariates X include the age, sex, race, baseline weight, and smoking intensity of each individual.

Of the selected 1566 individuals, 1163 are non-quitters and the other 403 are quitters. The mean weight gain of non-quitters and quitters is 1.98 kg and 4.53 kg respectively (see Fig. 4), which means that for the studied individuals, quitters experience approximately 2.55 kg more weight gain than non-quitters on average. However, as we have discussed, this associational mean difference $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$ is not the causal effect $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ of quitting smoking because of the existence of confounding bias. For example, in the observational data, older people are more likely to quit smoking and gain less weight than younger people regardless of whether they quit smoking. Moreover, more males quit smoking than females. As a result, naive estimation from the observational data may underestimate the true treatment effects and we need to adjust for these possible confounders for the purpose of treatment effect estimation. Following Hernan and Robins [8], we assume unconfoundedness conditional on the observed covariates.

We use the same Logistic regression model used in [8] to fit the treatment propensities and replicate the estimation 1000 times by randomly sampling 70% individuals from each group (quitter and non-quitter) in each replication. The estimated ATE and corresponding empirical standard errors are listed in Table 5.

[2] https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx/ .

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

Table 5
Estimation results for the NHEFS dataset. The ATE and standard errors (SE) are computed from 1000 replications. The reference estimation in [8] is 3.4 kg with a 95% confidence interval of 2.4 ~ 4.5 kg.

	ATE	SE
Naive	2.534	(0.010)
IPW	3.412	(0.010)
Trunc	2.466	(0.009)
TruncNorm	3.245	(0.010)
TruncCrump	3.354	(0.010)
Norm	3.432	(0.010)
PS	3.412	(0.010)
PSNorm	3.438	(0.010)

As we can see from the result, except for the *Naive* estimator, the *Trunc* estimator and the *TruncNorm* estimator, all the other estimators have a similar estimated ATE of about 3.4 kg, i.e., quitting smoking increases weight by about 3.4 kg for the investigated population. This estimate is quite close to that in [8], which is 3.4 kg with a 95% confidence interval of 2.4 ~ 4.5 kg. Though the ground truth is unknown, we can conclude from the result that the estimates of the proposed Pareto-smoothed estimators match existing unbiased estimators (the IPW estimator and the self-normalized IPW estimator). In addition, by comparing the estimates of different truncated IPW estimators, we find that the estimate of truncated estimators is sensitive to the selected truncation thresholds.

7. Conclusion and further study

In this paper, we reframe the classic IPW estimator for causal inference into the framework of expectation estimation using importance sampling. To handle extreme importance weights commonly existed in importance weight-based estimators using finite samples, we take advantage of the smoothing property of the GPD from the extreme value statistics and propose a new Pareto-smoothing method to stabilize the IPW causal estimator. Based on this method, we further propose two Pareto-smoothed causal estimators, the Pareto-smoothed IPW estimator and the Pareto-smoothed self-normalized IPW estimator. Comprehensive experiments using both simulated and semi-simulated data indicate that, for causal inference from finite observational data, the proposed Pareto-smoothed estimators generally achieve lower bias than estimators using weight truncation or weight self-normalization. Moreover, they are more stable than the vanilla IPW estimator and the self-normalized IPW estimator. We also validate the proposed method with a real-world health dataset.

Note that although we focus on IPW-based estimation of the ATE in this paper, the key component of the proposed method is in principle to stabilize the estimated importance weights by fitting a GPD over the tail to smooth the extreme weights. This is quite general and can be easily adapted for the estimation of other causal estimands (e.g., ATT and ATC) with any other propensity score based causal estimators. As a result, one of our future research undertakings will be to investigate the application of the proposed method in other causal estimators such as propensity score matching and weighted outcome regression.

In addition, we assume unconfoundedness and estimate the treatment propensities with all the observed pre-treatment covariates for simplicity in this paper. Many researchers have recently noticed that variable selection in propensity score estimation using the outcome adaptive LASSO [42] or the highly adaptive LASSO [43] can also stabilize the resulting propensity score-based estimators. We believe this will also be beneficial for our proposed estimators and leave that item for future study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Australian Research Council (ARC) under Discovery Grant DP170101632. The first author is sponsored by the China Scholarship Council (CSC No. 201506030111). We would like to thank Yiliao Song for discussion and the anonymous reviewers for their valuable comments. We also thank Susan Felix and Camila Cremonese for proofreading the paper.

Appendix A. Estimation of ATT

In the main text, we demonstrate the proposed method by focusing on the estimation of ATE. If the estimated of interest is the ATE for the treated, $\tau_{ATT} = \mathbb{E}[Y(1) - Y(0)|T = 1]$, the covariate distribution for our target population is then $p_X^{t=1} := p(X|T = 1)$. The expected potential outcomes $\mathbb{E}[Y(0)|T = 1]$ and $\mathbb{E}[Y(1)|T = 1]$ are estimated using importance sampling via

$$\mathbb{E}[Y(1)|T=1] = \mathbb{E}_{p_X^{t=1}}\mathbb{E}[Y|X] = \frac{1}{n_1}\sum_{i:T_i=1} Y_i$$

$$\begin{aligned}\mathbb{E}[Y(0)|T=1] &= \mathbb{E}_{p_X^{t=1}}\mathbb{E}[Y(0)|X] \\ &= \mathbb{E}_{p_X^{t=1}}\mathbb{E}[Y|X,T=0] \\ &= \frac{1}{n_0}\sum_{i:T_i=0}\frac{p(X_i|T_i=1)}{p(X_i|T_i=0)}Y_i \\ &= \frac{1}{n_0}\sum_{i:T_i=0}\frac{e(X_i)}{1-e(X_i)}Y_i\end{aligned}$$

Apparently, we only need to weight the individuals in the control group to match the treatment group. Define the importance weight for X_i as

$$W_i = \begin{cases} 1, & \text{if } T_i = 1 \\ \dfrac{e(X_i)}{1-e(X_i)}, & \text{if } T_i = 0 \end{cases} \tag{A.1}$$

After Pareto-smoothing W_i for individuals in the control group, we obtain the Pareto-smoothed importance weights $\{W_1^{PS}, W_2^{PS}, \ldots, W_n^{PS}\}$, and the Pareto-smoothed IPW estimator for ATT is defined as

Algorithm 2 Pareto-smoothed IPW ATT Estimator.

Input: Observation data $\mathcal{D} = \{(X_i, T_i, Y_i), \ldots, (X_n, T_n, Y_n)\}$
Output: The estimated $\hat{\tau}_{ATT}^{\text{PS}}$

1: Fit a treatment propensity model $e(X)$ form $\mathcal{D}$;
2: Calculate the importance weights for each unit by Eq. (A.1) to obtain $\{W_i, i = 1, \ldots, n\}$
3: Sort the importance weights $\{W_i, i = 1, \ldots, n\}$ ascendingly to obtain the sorted importance weights $W_{[1]}, W_{[2]}, \ldots, W_{[n]}$
4: Choose the location parameter $\hat{\mu}$ by Eq. (18))
5: Estimate the parameters σ and k by Eq. (23))
6: Smooth the importance weights $\{W_1, W_2, \ldots, W_n\}$ by Eq. (25)) to obtain the Pareto-smoothed importance weights $\{W_1^{\text{PS}}, W_2^{\text{PS}}, \ldots, W_n^{\text{PS}}\}$
7: Estimate the ATE $\hat{\tau}_{ATT}^{\text{PS}}$ via Eq. (A.2)

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

$$\hat{\tau}_{ATT}^{PS} = \frac{1}{n_1}\sum_{i:T_i=1} W_i^{PS} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} W_i^{PS} Y_i$$
$$= \frac{1}{n_1}\sum_{i:T_i=1} Y_i - \frac{1}{n_0}\sum_{i:T_i=0} W_i^{PS} Y_i \quad (A.2)$$

The implementation procedure is summarized in Algorithm 2.

Moreover, in the case of estimating the ATT, the corresponding estimation bias is defined as

$$Bias_{\mathrm{ATT}} = |\hat{\tau}_{ATT} - \tau_{ATT}| = \left|\hat{\tau}_{ATT} - \frac{1}{n_1}\sum_{t_i=1}(Y_i(1) - Y_i(0))\right|$$

References

[1] S. Shalev-Shwartz, S. Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014.
[2] S. Athey, Beyond prediction: using big data for policy problems, Science 355 (6324) (2017) 483–485.
[3] J. Pearl, Theoretical impediments to machine learning with seven sparks from the causal revolution, arXiv:1801.04016 (2018).
[4] S. Zhao, T. Liu, S. Zhao, Y. Chen, J.-Y. Nie, Event causality extraction based on connectives analysis, Neurocomputing 173 (2016) 1943–1950.
[5] S. Du, G. Song, H. Hong, Collective causal inference with lag estimation, Neurocomputing 323 (2019) 299–310.
[6] H. Zenil, N.A. Kiani, A.A. Zea, J. Tegnér, Causal deconvolution by algorithmic generative models, Nat. Mach. Intell. 1 (1) (2019) 58.
[7] G.W. Imbens, D.B. Rubin, Causal Inference in Statistics, Social, and Biomedical Sciences, Cambridge University Press, 2015.
[8] M.A. Hernan, J.M. Robins, Causal Inference, Chapman & HallCRC, Boca Raton, 2019. forthcoming
[9] B.M. Lake, T.D. Ullman, J.B. Tenenbaum, S.J. Gershman, Building machines that learn and think like people, Behav. Brain Sci. 40 (2017).
[10] F. Zhu, G. Zhang, J. Lu, D. Zhu, First-order causal process for causal modelling with instantaneous and cross-temporal relations, in: 2017 International Joint Conference on Neural Networks (IJCNN), IEEE, 2017, pp. 380–387.
[11] L. Ma, J. Dong, K. Peng, Root cause diagnosis of quality-related faults in industrial multimode processes using robust gaussian mixture model and transfer entropy, Neurocomputing 285 (2018) 60–73.
[12] S. Athey, G. Imbens, T. Pham, S. Wager, Estimating average treatment effects: supplementary analyses and remaining challenges, Am. Econ. Rev. 107 (5) (2017) 278–281.
[13] F. Zhu, A. Lin, G. Zhang, J. Lu, Counterfactual inference with hidden confounders using implicit generative models, in: Australasian Joint Conference on Artificial Intelligence, Springer, 2018, pp. 519–530.
[14] P.R. Rosenbaum, D.B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.
[15] F. Zhu, A. Lin, G. Zhang, J. Lu, D. Zhu, Pareto-smoothed inverse propensity weighing for causal inference, in: Data Science and Knowledge Engineering for Sensing Decision Support: Proceedings of the 13th International FLINS Conference (FLINS 2018), vol. 11, World Scientific, 2018, p. 413.
[16] D.B. Rubin, Causal inference using potential outcomes: design, modeling, decisions, J. Am. Stat. Assoc. 100 (469) (2005) 322–331.
[17] S.R. Seaman, I.R. White, Review of inverse probability weighting for dealing with missing data, Stat. Methods Med. Res. 22 (3) (2013) 278–295.
[18] D. Liang, L. Charlin, D.M. Blei, Causal inference for recommendation, Causation: Foundation to Application, Workshop at UAI, AUAI, 2016.
[19] T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, T. Joachims, Recommendations as treatments: debiasing learning and evaluation, in: International Conference on Machine Learning, 2016, pp. 1670–1679.
[20] X. Wang, M. Bendersky, D. Metzler, M. Najork, Learning to rank with selection bias in personal search, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2016, pp. 115–124.
[21] T. Joachims, A. Swaminathan, T. Schnabel, Unbiased learning-to-rank with biased feedback, in: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, ACM, 2017, pp. 781–789.
[22] A.B. Owen, Monte Carlo Theory, Methods and Examples (book draft), 2014.
[23] S. Coles, J. Bawa, L. Trenner, P. Dorazio, An Introduction to Statistical Modeling of Extreme Values, vol. 208, Springer, 2001.
[24] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, vol. 793, Wiley, 2019.
[25] J. Pearl, Causality: Models, Reasoning and Inference, Cambridge University Press, 2009.
[26] P.R. Rosenbaum, Model-based direct adjustment, J. Am. Stat. Assoc. 82 (398) (1987) 387–394.
[27] E.L. Ionides, Truncated importance sampling, J. Comput. Graph. Stat. 17 (2) (2008) 295–311.
[28] R.K. Crump, V.J. Hotz, G.W. Imbens, O.A. Mitnik, Dealing with limited overlap in estimation of average treatment effects, Biometrika 96 (1) (2009) 187–199.
[29] S. Yang, P. Ding, Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores, Biometrika 105 (2) (2018) 487–493.
[30] C. Ju, J. Schwab, M.J. van der Laan, On adaptive propensity score truncation in causal inference, Stat. Methods Med. Res. (2018). 0962280218774817
[31] M.J. Van der Laan, S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, Springer Science & Business Media, 2011.
[32] M.S. Schuler, S. Rose, Targeted maximum likelihood estimation for causal inference in observational studies, Am. J. Epidemiol. 185 (1) (2017) 65–73.
[33] J. Pickands III, et al., Statistical inference using extreme order statistics, Ann. Stat. 3 (1) (1975) 119–131.
[34] A. Vehtari, A. Gelman, J. Gabry, Pareto smoothed importance sampling, arXiv:1507.02646 (2015).
[35] A. Vehtari, A. Gelman, J. Gabry, Practical Bayesian model evaluation using leave-one-out cross-validation and Waic, Stat. Comput. 27 (5) (2017) 1413–1432.
[36] E.B. Mackay, P.G. Challenor, A.S. Bahaj, A comparison of estimators for the generalised Pareto distribution, Ocean Eng. 38 (11–12) (2011) 1338–1346.
[37] J. Zhang, M.A. Stephens, A new and efficient estimation method for the generalized Pareto distribution, Technometrics 51 (3) (2009) 316–325.
[38] R.C. Nethery, F. Mealli, F. Dominici, Estimating population average causal effects in the presence of non-overlap: a Bayesian approach, arXiv:1805.09736 (2018).
[39] J.L. Hill, Bayesian nonparametric modeling for causal inference, J. Comput. Graph. Stat. 20 (1) (2011) 217–240.
[40] U. Shalit, F.D. Johansson, D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3076–3085.
[41] C. Louizos, U. Shalit, J.M. Mooij, D. Sontag, R. Zemel, M. Welling, Causal effect inference with deep latent-variable models, in: Advances in Neural Information Processing Systems, 2017, pp. 6446–6456.
[42] S.M. Shortreed, A. Ertefaie, Outcome-adaptive lasso: variable selection for causal inference, Biometrics 73 (4) (2017) 1111–1122.
[43] C. Ju, D. Benkeser, M.J. van der Laan, Flexible collaborative estimation of the average causal effect of a treatment using the outcome-highly-adaptive lasso, arXiv:1806.06784 (2018).

Fujin Zhu is currently pursuing dual Ph.D. degrees with the School of Management and Economics, Beijing Institute of Technology (BIT), Beijing, China and Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. He received his B.E. degree in Management Science and Engineering from BIT in 2013. His research interests includes causal inference and machine learning. He is a Member of the Decision Systems and eService Intelligence (DeSI) Research Laboratory, CAI, University of Technology Sydney.

Jie Lu is a Distinguished Professor and the Director of the Centre for Artificial Intelligence at the University of Technology Sydney, Australia. She received her PhD degree from Curtin University of Technology, Australia, in 2000. Her main research expertise is in fuzzy transfer learning, decision support systems, concept drift, and recommender systems. She has published six research books and 400 papers in Artificial Intelligence, IEEE Transactions on Fuzzy Systems and other refereed journals and conference proceedings. She has won over 20 Australian Research Council (ARC) discovery grants and other research grants worth more than $4 million. She serves as Editor-In-Chief for Knowledge- Based Systems (Elsevier) and Editor-In-Chief for International Journal on Computational Intelligence Systems (Atlantis), has delivered 20 keynote speeches at international conferences, and has chaired 10 international conferences. She is a Fellow of IEEE and Fellow of IFSA.

Adi Lin is currently pursuing the Ph.D. degree with the Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. He received his B.S. degree in computer science and the M.S. degree in computer software and theory from the School of Information Science and Engineering, Xiamen University, Xiamen, China, in 2012 and 2015, respectively. His research interests includes causal inference and Bayesian Nonparametric models. He is a Member of the Decision Systems and eService Intelligence (DeSI) Research Laboratory, CAI, University of Technology Sydney.

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095

Guangquan Zhang is a professor and Director of the Decision Systems and e- Service Intelligent (DeSI) Research Laboratory, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. He received his PhD in applied mathematics from Curtin University of Technology, Australia, in 2001. His research interests include fuzzy machine learning, fuzzy optimization, and machine learning and data analytics. He has authored four monographs, five textbooks, and 350 papers including 160 refereed international journal papers. Dr. Zhang has won seven Australian Research Council (ARC) Discovery Project grants and many other research grants. He was awarded an ARC QEII Fellowship in 2005. He has served as a member of the editorial boards of several international journals, as a guest editor of eight special issues for IEEE Transactions and other international journals, and has cochaired several international conferences and workshops in the area of fuzzy decision-making and knowledge engineering.

Please cite this article as: F. Zhu, J. Lu and A. Lin et al., A Pareto-smoothing method for causal inference using generalized Pareto distribution, Neurocomputing, https://doi.org/10.1016/j.neucom.2019.09.095