Survey on Causal Inference and Heterogeneous Treatment Effects

Qiming Du

February 11, 2020

Abstract

In this shallow document we collect some basic materials on the Causal Inference and its applications on the estimation of Heterogeneous Treatment Effects (HTE). The topics cover the basic setting of the Estimation of HTE, i.e., Average Treatment Effect (ATE), Conditional Average Treatment Effect (CATE) or Individual Treatment Effect (ITE) and unconfoundedness assumption, etc. We also discuss the basic meta-learners: T-learners, S-learners and X-learners, and R-learners.

The topics will be updated when I learn new things in this domain. At the moment, the main references are:

- Survey on Causal Inference: [Pea09];
- Meta-learners (S-,T- and X-learners): [KSBY17];
- Causal Forests: [AW19];
- R-learners: [NW17].

Contents

1 General background on Causal Inference

Before starting, we give some conceptional discussions on Causal Inference. Our goal is to briefly summary some main ideas introduced in [Pea09] and to introduce some of the most important vocabularies from a statistical perspective. Hence, we do not emphasize on the existing methods and models. The details can be found in [Pea09] and references therein.

Roughly speaking, by "causality" we mean that the change/modification of a random variable X can affect another random variable Y. The "slogan" of Causal Inference can be stated as follows:

Joint distribution does not contain the information on causality.

That being said, no probability modeling is capable of calculating causality without further assumptions. In order to take advantage of statistical inference, the solution at the moment is direct and natural: we make assumptions that are not testable by observed data, and under such assumptions, Bayesian Inference can be exploited to provide Causal Inference. There are mainly two existing ways to describe the causal assumptions. The calculus of

$$\mathbf{P}\left(Y \mid do(X = x_0)\right)$$

to represent *interventions* and Structural Equation Modelling (SEM).

Remark 1.1. I have made some assumptions in order to start in a relatively simple situation:

- I assume that the study of HTE is *always* (at least, for the first step) under potential outcome framework with unconfoundedness assumption and that the propensity score is bounded respectively by $0 < e_{min}$ and $e_{max} < 1$. Hence, I did not investigate the classical methods in Causal Inference such as instrumental variables, etc. It seems that when unconfoundedness is missing, one may consider using instrumental variables to conduct causal inference. (As mentioned in page 3 [NW17])
- I did not investigate influence of the censored data. I imagine that this is a separate problem that can be addressed later by modifying the regressors we used.

I still do not fully understand the difference between *randomized study* and *observational study* (briefly mentioned in the beginning of [AW19]). It seems to me that the difference is on the treatment assignments, i.e., in the randomized study, the treatment assignment indicator is independent to *any* random variables, while in the observational study, it may depend on the covariates. Any comments? What can not be done in the observational study?

2 Potential outcome framework

In this section, we present the classic Neyman-Rubin potential outcome framework (see, e.g., [Rub74] and [IR15]) widely used in the study of HTE. We assume that the data consist in N i.i.d. samples of triplets

$$(Y_i(W_i), W_i, X_i)$$

where

- $X_i \in \mathbf{R}^d$ is a d-dimensional covariate or feature vector;
- $W_i \in \{0, 1\}$ denotes the treatment assignment indicator;
- $Y_i(1)$ and $Y_i(0)$ denote respectively the potential outcome of individual i with and without treatment. With a slight abuse of notation, we denote $Y_i := Y_i(W_i)$.

The canonical random variable is denoted by

$$(Y(0), Y(1), W, X).$$

We denote $\mathcal{X}$ the marginal distribution of X, and $\mu_\omega(x) := \mathbf{E}\left[Y(\omega) \mid X = x\right]$ for $\omega \in \{0, 1\}$. The *propensity score* is defined by $e(x) := \mathbf{P}\left(W = 1 \mid X = x\right)$.

Assumption 1 (Unconfoundedness, [RR83])**.** *We assume that the treatment assignment is as good as random conditionally on covariates, namely,*

$$\forall i \in [N], \quad (Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i \ ,$$

or equivalently,

$$(Y(0), Y(1)) \perp\!\!\!\perp W \mid X.$$

Assumption 2 (Well-posedness or Overlap). *We assume that e_{min} and e_{max} exist such that the inequality*

$$0 < e_{min} < e(x) < e_{max} < 1$$

holds uniformly for all x in the support of X.

For the existing methods, the main goal is to estimate the *Conditional Average Treatment Effect* (CATE) function defined by

$$\tau(x) := \mathbf{E}\left[Y(1) - Y(0) \mid X = x\right].$$

This allows to conduct causal inference on a new individual with covariate x_j. Therefore, for each individual x_j, we also call

$$D_j := Y_j(1) - Y_j(0) = \tau(x_j)$$

the *Individual Treatment Effect* (ITE).

(In [KSBY17], there is an argument that writes: the best estimator for CATE is also the best estimator for ITE w.r.t. MSE. Personally, I do not see the meaning of introducing such a notation which is quite misleading for not emphasizing the randomness of the new comers. It provides essentially the same information as by CATE function. Therefore, I try to avoid the usage of ITE in the following. If everyone agrees, I erase this paragraph in the following week. Tell me if I missed something important.)

3 Existing methods

In this section, we present some existing strategy used to estimate the CATE function. At the moment, they are divided into three classes: Meta-learners, Causal Forests and R-learners.

3.1 Meta-learners

According to the authors of [KSBY17], the Meta-learners framework aims at providing a general strategy where the CATE function estimation problem in potential outcome framework can be divided into several supervised learning problems. At the same time, there is no need to modify the existing supervised learning algorithms so that they can be implemented. This is one of the main differences of the philosophy behind the design of Meta-learners and Causal Forests [AW19] as well as R-learners [NW17], where the latter require an adaptation of the existing regression models. Before proceeding further, we introduce some notation which are useful in the following, especially in the Meta-learners context. A model trained by the training data $(X_i, Y_i; 1 \leq i \leq N)$ is denoted by

$$\mathcal{M}_{\{X_i \sim Y_i \mid i \in [N]\}}.$$

The associated prediction of the *conditional expectation* at point x is denoted by

$$\mathbf{Pred}\left(x \mid \mathcal{M}_{\{X_i \sim Y_i \mid i \in [N]\}}\right).$$

This notation is more compatible with the models in the Bayesian nonparametrics context, emphasizing the randomness brought by the model itself. When the underlying model is Random Forests-based, we replace $\mathcal{M}$ by $\mathcal{T}$. Below we list the construction of T-, S- and X-learners.

- T-learners:
  Divided by two subgroups w.r.t. W_i, we consider the estimator $\hat{\mu}^T_\omega$ of μ_ω defined as follows:

  $$\hat{\mu}^T_\omega(x) := \mathbf{Pred}\left(x \mid \mathcal{M}_{\{X_i \sim Y_i \mid W_i = \omega,\ i \in [N]\}}\right).$$

  Then, the estimator $\hat{\tau}^T$ of τ is defined by

  $$\hat{\tau}^T(x) = \hat{\mu}^T_1(x) - \hat{\mu}^T_0(x).$$

  (With a slight abuse of notation, we may denote $\hat{\mu}_\omega := \hat{\mu}^T_\omega$.)

- S-learners:
  Without specifying the role of W_i, the estimator $\hat{\mu}_\omega$ of μ_ω defined as follows:

  $$\hat{\mu}^S_\omega(x) := \mathbf{Pred}\left((x, \omega) \mid \mathcal{M}_{\{(X_i, W_i) \sim Y_i \mid i \in [N]\}}\right).$$

  Then, the estimator $\hat{\tau}^S$ of τ is defined by

  $$\hat{\tau}^S(x) = \hat{\mu}^S_1(x) - \hat{\mu}^S_0(x).$$

- X-learners:

  By estimating the "missing" potential outcomes with the same spirit of T-learners, we define

$$\hat{\tau}_1^X(x) := \mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim Y_i - \hat{\mu}_0^T(X_i) \mid W_i=1,\ i\in[N]\}}\right),$$

  and

$$\hat{\tau}_0^X(x) := \mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim \hat{\mu}_1^T(X_i) - Y_i \mid W_i=0,\ i\in[N]\}}\right).$$

  The final estimator $\hat{\tau}^X$ of τ is constructed by taking the convex combination of the two estimators defined above, that is

$$\hat{\tau}^X(x) := g(x)\hat{\tau}_1^X(x) + (1 - g(x))\hat{\tau}_0^X(x).$$

  For the choice of $g(x)$, one may consider the estimator of the propensity score $\hat{e}(x)$, or one can simply take 0 or 1 according to the data.

3.2 Causal Forests

Causal Forests can be regarded as an adaption of the classic Random Forests to HTE estimation problems. More precisely, the splitting criterion is modified, which optimize the same objective (inspired by [CCD+18] for the case where τ is a constant function) of R-learners [NW17]. For each tree, when the partition is made, the prediction is simply the average of the outcome differences in the same leaf. The modification of the splitting enters into the new Generalized Random Forests framework introduced in [ATW19]. In a high level, Causal Forests can be seen as a Random Forests-based method such that the split w.r.t. W_i always occurs in the last step.

3.3 R-learners

TODO.

(The construction of R-learners is from a totally different perspective from Meta-learners and Causal Forests. We will discuss it later.)

4 Uncertainty control of CATE function

The traditional literature on HTE estimation concentrate on the estimation of *CATE* function $\tau(x)$. In this section, we discuss the possibility of going one step further, i.e., to control the uncertainty of the CATE function. More precisely, we are interested in estimating

$$\mathbf{P}\left(|(Y(1) - Y(0)) - \tau(x)| > a \mid X = x\right). \tag{1}$$

4.1 Partial results in general case

Unfortunately, since the information on the dependence of the conditional distributions $\mathcal{L}(Y(0) \mid X = x)$ and $\mathcal{L}(Y(1) \mid X = x)$ is required to perform the estimation of (1), which may be rarely collected by the dataset, we consider an upper bound of the probability given in (1) w.r.t. a. Before proceeding further, we introduce some additional notation to simplify the writing. We denote

$$\mu_{00}(x) := \mathbf{E}\left[Y(0)^2 \,\middle|\, X = x\right],$$

and

$$\mu_{11}(x) := \mathbf{E}\left[Y(1)^2 \,\middle|\, X = x\right],$$

as well as

$$\mu_{01}(x) := \mathbf{E}\left[Y(0)Y(1) \mid X = x\right].$$

The conditional variances are denoted respectively by

$$\begin{aligned}\sigma_{00}(x) :=& \mathbf{E}\left[Y(0)^2 \,\middle|\, X = x\right] - \mathbf{E}\left[Y(0) \mid X = x\right]^2, \\ =& \mu_{00}(x) - \mu_0(x)^2,\end{aligned}$$

and

$$\begin{aligned}\sigma_{11}(x) :=& \mathbf{E}\left[Y(1)^2 \,\middle|\, X = x\right] - \mathbf{E}\left[Y(1) \mid X = x\right]^2, \\ =& \mu_{11}(x) - \mu_1(x)^2.\end{aligned}$$

Thanks to Chebyshev's inequality, we have

$$\begin{aligned}
&\mathbf{P}\left(|(Y(1)-Y(0))-\tau(x)|>a \mid X=x\right)\\
\leq &\frac{\operatorname{Var}\left[Y(1)-Y(0) \mid X=x\right]}{a^2}\\
= &\frac{\sigma_{00}(x)+\sigma_{11}(x)-2\mu_{01}(x)+2\mu_0(x)\mu_1(x)}{a^2}.
\end{aligned} \tag{2}$$

Moreover, Cauchy-Schwarz inequality gives

$$|\mu_{01}(x)| \leq \sqrt{\mu_{00}(x)\mu_{11}(x)},$$

whence

$$\begin{aligned}
&\mathbf{P}\left(|(Y(1)-Y(0))-\tau(x)|>a \mid X=x\right)\\
\leq &\frac{\sigma_{00}(x)+\sigma_{11}(x)+2\mu_0(x)\mu_1(x)+2\sqrt{\mu_{00}(x)\mu_{11}(x)}}{a^2}.
\end{aligned} \tag{3}$$

The interests of the upper bound given in (3) lies in the fact that all the terms at the r.h.s. can be approximated with supervised learning algorithm, under certain regularity assumptions. In practice, this allows a finer analysis on the causal inference of the treatment. Some numerical illustration of the estimation of conditional variance can be found in the following notebook:

- https://mgimm.github.io/doc/jupyternb/Estimation-of-Conditional-Variance.html

Another statistics that may be interesting for the causal inference can be defined by

$$\ell_a(\beta) := \arg\min_{a\in\Pi}\left\{\frac{\sigma_{00}(x)+\sigma_{11}(x)+2\mu_0(x)\mu_1(x)+2\sqrt{\mu_{00}(x)\mu_{11}(x)}}{a^2} \leq \beta\right\},$$

where Π is some prefixed discretization of $\mathbf{R}_+$, so that "arg min" is well-defined. Equivalently, this statistics measures (an upper bound of) the standard variation of the conditional distribution $\mathcal{L}\left(Y(1)-Y(0) \mid X=x\right)$.

4.2 Uniform noise

In order to simplify the study of the dependence of $\mathcal{L}\left(Y(0) \mid X=x\right)$ and $\mathcal{L}\left(Y(1) \mid X=x\right)$, we suppose that for all $x \in E$, we have

$$Y(0) = \mu_0(X)+\epsilon_0,$$

and

$$Y(1) = \mu_1(X)+\epsilon_1,$$

where the distributions of ϵ_0 and ϵ_1 do not depend on X. Note that this is the same setting considered in [KSBY17], where the dependence of $\mathcal{L}\left(Y(0) \mid X=x\right)$ and $\mathcal{L}\left(Y(1) \mid X=x\right)$ has nothing to do with the value of X. In this case, if we suppose the consistence of the estimators $\hat{\mu}_0$ and $\hat{\mu}_1$, one can deduce the i.i.d. samples of ϵ_0 and ϵ_1 by considering

$$\widetilde{Y}_i(0) := Y_i(0)-\hat{\mu}_0(X_i) \quad \text{for } W_i = 0,$$

and

$$\widetilde{Y}_i(1) := Y_i(1)-\hat{\mu}_1(X_i) \quad \text{for } W_i = 1.$$

The estimation of the joint distribution of (ϵ_0, ϵ_1) can thus be derived by the samples $(Y_i(0); W_i = 0, 1 \leq i \leq N)$ and $(Y_i(1); W_i = 1, 1 \leq i \leq N)$. In this case, the conditional confidential interval defined in (1) can be evaluated/approximated directly supposing the consistence of some subproblems of supervised learning. (to be discussed)

5 Wasserstein Random Forests

In this section, we discuss the Forests-based conditional distribution estimation strategy, which aims to solving a subproblem of HTE estimation (cf. Remark A.4).

5.1 Setting

We consider a canonical $(d+q)$-dimensional random variable (X, Y), where X is a d-dimensional covariate. The support of X is denoted by $E \subset \mathbf{R}^d$. The goal is to estimate the conditional distribution $\mathcal{L}(Y \mid X = x)$. The associated measure is denoted by $\pi_x(dy)$. The Wasserstein distance $\mathcal{W}_p$ between two measures μ, ν on $\mathbf{R}^q$ is defined by

$$\mathcal{W}_p(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int |x-y|^p \, \gamma(dx, dy) \right)^{\frac{1}{p}},$$

where $\Gamma(\mu, \nu)$ denotes the set of all couplings of μ and ν, namely,

$$\int_{y \in \mathbf{R}^q} \gamma(dx, dy) = \mu(dx) \quad \text{and} \quad \int_{x \in \mathbf{R}^q} \gamma(dx, dy) = \nu(dy).$$

Now, let us introduce some regularity assumptions.

Assumption 3 (Lipschitz). *For any $(x, y) \in E \times E$, we assume that there exists a constant $C_\pi < +\infty$ such that*

$$\mathcal{W}_2(\pi_x, \pi_y) \leq C_\pi \, |x-y| \, .$$

Assumption 4 (Hölder). *For any $(x, y) \in E \times E$, we assume that there exists $\alpha > 0$ and a constant $C_\pi^\alpha < +\infty$ such that*

$$\mathcal{W}_2(\pi_x, \pi_y) \leq C_\pi^\alpha \, |x-y|^\alpha \, .$$

I assume that these two assumptions are both standard in this setting.

Our goal is to construct an estimator π_x^N of π_x such that for all $x \in E$, we have

$$W_2(\pi_x^N, \pi_x) \xrightarrow[N \to \infty]{} 0.$$

Note that the convergence above is equivalent to

$$\pi_x^N \xrightarrow[N \to \infty]{\mathrm{d}} \pi_x$$

and the convergence of first and second moments. In particular, the convergence w.r.t. $\mathcal{W}_2$-distance ensures the convergence of the naive conditional variance estimator $\pi_x^N(y \mapsto y^2) - \pi_x^N(y \mapsto y)^2$ if $\sup_{x \in E} \pi_x(\mathrm{Id}) < +\infty$.

5.2 Splitting criterion

We present an intuitive idea on how to perform the splitting which is compatible with the Wasserstein distance. Roughly speaking, we try to maximize the $\mathcal{W}_2$-distance between the empirical measures w.r.t. Y_i of the newly created left node and right node, which obeys the same philosophy of the classical Random Forests: the splitting is performed greedily in order to maximize the homogeneity between the subgroups. This criterion also works when the dimension of Y is greater than 1. However, in this case, the calculation of $\mathcal{W}_2$-distance may becomes intractable. Some regularization strategies (such as sinkhorn distance, i.e., OT with entropic constraints) may be applied.

Appendices

A Minutes

A.1 31 January 2020: FP,QD,RP

A.1.1 Numerical implementations

There is an extremely well-documented tools available on github, provided by **Uber**:

- https://github.com/uber/causalml.

A more detailed notebook can be found here:

- https://github.com/uber/causalml/blob/master/examples/meta_learners_with_synthetic_data.ipynb

From what I have tested, in the notebook of the Meta-learners above, there is a problem when implementing the training procedure without providing an estimation of propensity score. (e.g., In[8] and In[10] do not work.) A quick solution is that we estimate the propensity score function separatly. I will look into this problem this week. Other than that, this package is well-written: it provideds a generator of synthetic data with different modes; the choice of base-learners is totally free; the summary tools are well designed; there is also some feature interpretation strategy that I have not studied yet:

- https://github.com/uber/causalml/blob/master/examples/feature_interpretations_example.ipynb.

To conclude, if we want to test Meta-learners and R-learners with our own data, there are already some quite good tools to begin with. Other than this, I plan to write some codes for X-learners this week, and I will explain why in the next section.

A.1.2 New ideas

The existing methods focus on the estimation problem of the CATE function $\tau(x)$. However, the information we can get from $\tau(x)$ is limited. Instead, we are interested in estimating the conditional law:

$$\mathcal{L}_Y(x) := \mathcal{L}\left((Y(1), Y(0)) \mid X = x\right). \tag{4}$$

We also denote $\pi_x(dy)$ the associated probability measure. With $\mathcal{L}_Y(x)$ we can do some far more complicated inference on the behaviors of $Y(0)$ and $Y(1)$ given the covariate of a new individual x. In order to achieve this, we need some assumptions. I divide this problem into 3 stages.

Stage 1: parametric models We start with the most simple setting: we assume that $\mathcal{L}_Y(x)$ is a 2-dimensional Gaussian random variable, i.e.,

$$\mathcal{L}_Y(x) = \mathcal{N}(m(x), \Sigma(x)).$$

Though it is a quite naive strategy, it can be regarded as a natural generalization of X-learners from a different point of view. In fact, in the context of X-learners, we only estimate the expectation $m(x)$. We can use the same idea to estimate the covariance matrix $\Sigma(x)$. More precisely, we denote

$$\Sigma(x) = \begin{pmatrix} \sigma_{00}(x) & \sigma_{01}(x) \\ \sigma_{10}(x) & \sigma_{11}(x) \end{pmatrix}.$$

Accordingly, we consider the estimators

$$\hat{\sigma}_{00}(x) := \mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim Y_i^2 \mid W_i=0,\ i\in[N]\}}\right) - \hat{\mu}_0^T(x)^2,$$

and

$$\hat{\sigma}_{11}(x) := \mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim Y_i^2 \mid W_i=1,\ i\in[N]\}}\right) - \hat{\mu}_1^T(x)^2,$$

as well as

$$\begin{aligned} &\hat{\sigma}_{01}(x) = \hat{\sigma}_{10}(x) \\ :=& g(x)\mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim \hat{\mu}_0^T(X_i) Y_i \mid W_i=1,\ i\in[N]\}}\right) && \text{(This is wrong!)} \\ &+ (1-g(x))\mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim Y_i \hat{\mu}_1^T(X_i) \mid W_i=0,\ i\in[N]\}}\right) - \hat{\mu}_0^T(x)\hat{\mu}_1^T(x), \end{aligned}$$

where $g(x)$ can be any $[0, 1]$-valued function, which can be chosen as an estimator of the propensity score. As we can see, the construction of the estimators above is of same taste as in the X-learners context. Said differently, no modification of the supervised learning algorithm is needed in order to conduct the causal inference. In this way, we have a relatively rough estimation of the conditional joint distribution $\mathcal{L}_Y(x)$.

Remark A.1. The numerical tests show that the esimator $\hat{\sigma}_{01}$ is poorly designed. In fact, since it is impossible to have the potential outcomes $Y(0)$ and $Y(1)$ at the same time, it is in general very difficult to get the information on the dependence of $Y(0)$ and $Y(1)$. The idea of filling the missing values with another supervised learning model is wrong in general, as the model can only provide an estimation for the conditional expectation, which is not enough for estimation of the covariance.

Stage 2: non-parametric models The basic idea is the similar as presented in [PL19], namely, to combine Random Forests and Kernel Density Estimation methods. This idea provide a better understanding on how the uncertainty propagates when the tree-learners are implemented. The idea of using Random Forests to estimate conditional distribution is also studied in the form of quantile regression (see, e.g., [Mei06]).

We will explain it in a relatively simple 1-dimensional toy example in the following. Let us assume that the kernel density function is Dirac mass δ_x. We skip the procedure of creating a tree (partition), and we assume the splitting criterion remains identical to the conditional expectation estimation problem. As is revealed by a lot of previous works (see, e.g., [AW19, Sco16]), the estimation of the Random Forests can be seen as some data-adaptive kernel methods. More precisely, we denote B the number of decision trees in a trained Random Forest, for each $b \in [B]$, we denote $L_b(x)$ the leaf that contains x. The subsample used to train the tree b is denoted by $\mathcal{S}_b \subset [N]$. Note that we assumed that bootstrap is not implemented for the construction of $\mathcal{S}_b$. For a random forest that contains B trees, the prediction at the point x is defined by

$$\begin{aligned}\textbf{Pred}\left(x \mid \mathcal{T}_{\{X_i \sim Y_i \mid i\in[N]\}}\right) &:= \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{N}\frac{Y_i \mathbf{1}_{\{X_i\in L_b(x) \mid i\in\mathcal{S}_b\}}}{\#\{i \mid X_i \in L_b(x), i \in \mathcal{S}_b\}}\\ &= \sum_{i=1}^{N} Y_i \underbrace{\frac{1}{B}\sum_{b=1}^{B}\frac{\mathbf{1}_{\{X_i\in L_b(x) \mid i\in\mathcal{S}_b\}}}{\#\{i \mid X_i \in L_b(x), i \in \mathcal{S}_b\}}}_{=:\alpha_i(x)}.\end{aligned}$$

By definition, we have

$$\forall x \in \textbf{Supp}(X), \quad \sum_{i=1}^{N}\alpha_i(x) = 1.$$

A very natural idea to approximate the conditional distribution

$$\pi_x(dy) = \mathcal{L}(Y \mid X = x)$$

is to take the weighted empirical measure

$$\hat{\pi}_x(dy) := \sum_{i=1}^{N}\alpha_i(x)\delta_{Y_i}.$$

Remark A.2. This idea is more or less identical to the quantile regression with Random Forests [Mei06]. However, we think the splitting criterion can be modified since our goal is indeed different from estimating the quantile. This kind of modification is also mentioned in [ATW19]. One possible idea will be discussed in Section 5.

More generally, when smoothing is needed, one may consider the kernel density function $K(\cdot)$ and the estimator:

$$\hat{\pi}_x^K(dy) := \sum_{i=1}^{N}\alpha_i(x)K\left(\frac{y - Y_i}{h}\right)dy,$$

with some window parameter h. When the training data is large, it is then not convenient to write down the density function first and conduct causal inference later. We provide another idea on how to implement numerically the inference, that is, instead of compute the density function, we sample from the distribution $\mathcal{L}(Y \mid X = x)$. When the kernel is set to be Dirac mass, the sampling is straightforward:

(i) Draw $A \sim \text{Categorical}\left(\alpha_1(x), \alpha_2(x), \dots, \alpha_N(x)\right)$;

(ii) Sample $Y_{new} \sim Y_A$.

The inference can thus be done by naive Monte Carlo methods. When the kernel density estimation is implemented, we consider

(i) Draw $A \sim \text{Categorical}\left(\alpha_1(x), \alpha_2(x), \dots, \alpha_N(x)\right)$;

(ii) Sample $Y_{new} \sim K\left(\frac{y-Y_A}{h}\right)dy$.

Remark A.3. There are several problems to be solved:

(i) The HTE estimation problem is essentially a 2-dimensional problem. Hence, the choice of kernel density function have to be taken into consideration. For example, if we choose Gaussian kernels, the covariance should be set to 0 automatically?

(According to the numerical tests, it is in general very difficult to get information on the dependence or covariance. Therefore, we first invertigate the 1-dimensional problem for the marginal distributions, namely, $\mathcal{L}(Y(0) \mid X = x)$ and $\mathcal{L}(Y(1) \mid X = x)$.)

(ii) The splitting strategy is difficult to design for a 2-dimensional distribution. Is W_2-distance for empirical measures (i.e., the empirical measures in the current leaf) a good direction to explore? Can this choice enters into GRF framework [AW19]? A possible and simple solution is to implement splitting which does not require the value of the objectives, such as Mondrian Forests [LRT14]. (The discussion of this topic can be found in Section 5.)

(iii) There is no theoretical guarantee, which I imagine is a quite difficult problem. (In fact, there is indeed theoretical guarantee for the quantile regression.)

Stage 3: robust non-parametric models Before starting, we would like to mention that by robustness we mean that there is very few specific tuning needed, as in the Random Forests context. I wonder if the *Pólya tree* can be used to estimate the conditional density. I am still reading and exploring in this direction. Any comments? other candidates? (This direction will be considered after the Forests-based conditional distribution estimation.)

Question A.1. *What can we do* exactly *if we have $\mathcal{L}_Y(x)$ available? Can we have a concrete example that can be used to compare with the inference* only *with CATE function?*

Remark A.4. According to RP, before considering more complicated inferences, such as to estimate

$$\mathbf{P}\left(Y(0) < a,\ Y(1) > b \mid Y(1) < c, X = x\right),$$

we could concentrate on the uncertainty control of the estimation of CATE function. This does not requires the information on the dependence of $(Y(0), Y(1))$. In fact, one only need to estimate the marginal distributions $\mathcal{L}\left(Y(0) \mid X = x\right)$ and $\mathcal{L}\left(Y(1) \mid X = x\right)$. This problem can be approached by considering the following separate subproblems, when this is no notable unbalance in treatment assignments.

- Estimate $\mathcal{L}\left(Y(0) \mid X = x\right)$ with $\{(X_i, Y_i(0)) \mid W_i = 0,\ i \in [N]\}$;
- Estimate $\mathcal{L}\left(Y(1) \mid X = x\right)$ with $\{(X_i, Y_i(1)) \mid W_i = 1,\ i \in [N]\}$.

This is identical to the logic behind the construction of the T-learners. The only difference is that the conditional expectations is replaced by conditional distributions as the objects to be approximated. If these two subproblems can be solved, a built-in “confidential interval” of the estimation of CATE function can also be obtained. The details are provided in Section 4.

A.2 07 February 2020: FP,QD,RP

A.2.1 Gaussian synthetic data

In order to start in a simple situation, we consider the following setting for the synthetic dataset:

- $X \sim \mathcal{N}(\mathbf{0}, \mathbf{Id})$ or Uniform($[0,1]^{\mathrm{d}}$);
- $(Y(0), Y(1)) \sim \mathcal{N}\left(m(X), \Sigma(X)\right)$ where

$$m(X) = (m(X)(0), m(X)(1)),$$

and

$$\Sigma(X) = \begin{pmatrix} \sigma_{00}(X) & \sigma_{01}(X) \\ \sigma_{01}(X) & \sigma_{11}(X) \end{pmatrix}.$$

- $W \sim$ Bernoulli($e(X)$).

A.2.2 Estimation of covariance matrix

The notebook of the numerical tests can be found here:
https://mgimm.github.io/doc/jupyternb/RDV-07-02-2020.html
Unfortunately, the quality for the estimation of covariance matrix is poor even in some extremely easy situation (e.g., when the covariance matrix is constant w.r.t. X).

Why it is still possible to estimate the variance function $\sigma_{00}(x)$ and $\sigma_{11}(x)$ We present some little calculations to show that why, under certain regularity assumptions, it is possible to estimate the conditional variances $\sigma_{00}(x)$ and $\sigma_{11}(x)$. For each $\omega \in \{0,1\}$, we denote

$$\hat{\mu}_\omega(x) = \mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim Y_i \mid W_i = \omega\}}\right),$$

and

$$\hat{\mu}_{\omega\omega}(x) = \mathbf{Pred}\left(x \,\middle|\, \mathcal{M}_{\{X_i \sim Y_i^2 \mid W_i = \omega\}}\right).$$

We assume that, for $\omega \in \{0,1\}$, we have

- $\hat{\mu}_\omega(x) - \mu_\omega(x) = \mathcal{O}_\mathbf{p}(a_N)$;
- $\hat{\mu}_{\omega\omega}(x) - \mu_{\omega\omega}(x) = \mathcal{O}_\mathbf{p}(b_N)$;

In addition, we also assume that

$$\|\hat{\mu}_\omega + \hat{\mu}_{\omega\omega}\|_\infty < +\infty. \quad a.s. \quad \text{(Alternatively, bounded in probability w.r.t. } x\text{?)}$$

Then, standard calculations assure that

$$\left(\hat{\mu}_{\omega\omega}(x) - \hat{\mu}^2_\omega(x)\right) - \underbrace{\left(\mu_{\omega\omega}(x) - \mu^2_\omega(x)\right)}_{\mathrm{Var}[Y(\omega)\mid X=x]} = \mathcal{O}_\mathbf{p}\left(a_N \vee b_N\right).$$

An implementation can be found here:

- https://mgimm.github.io/doc/jupyternb/Estimation-of-Conditional-Variance.html

B Useful links

- Missing data (Thanks Aude!)
  reference: https://rmisstastic.netlify.com/bibliography/
- XGBoost
  Tutorial: https://xgboost.readthedocs.io/en/latest/tutorials/index.html
  Parameters: https://xgboost.readthedocs.io/en/latest/parameter.html

References

[ATW19] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Ann. Statist.*, 47(2):1148–1178, 04 2019.

[AW19] Susan Athey and Stefan Wager. Estimating Treatment Effects with Causal Forests: An Application. *arXiv e-prints*, page arXiv:1902.07409, Feb 2019.

[CCD+18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015.

[KSBY17] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *arXiv e-prints*, page arXiv:1706.03461, Jun 2017.

[LRT14] Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3140–3148. Curran Associates, Inc., 2014.

[Mei06] Nicolai Meinshausen. Quantile regression forests. *J. Mach. Learn. Res.*, 7:983–999, 2006.

[NW17] Xinkun Nie and Stefan Wager. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *arXiv e-prints*, page arXiv:1712.04912, Dec 2017.

[Pea09] Judea Pearl. Causal inference in statistics: An overview. *Statist. Surv.*, 3:96–146, 2009.

[PL19] Taylor Pospisil and Ann B. Lee. (f)RFCDE: Random Forests for Conditional Density Estimation and Functional Data. *arXiv e-prints*, page arXiv:1906.07177, Jun 2019.

[RR83] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[Rub74] D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[Sco16] E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, March 2016.