Feature Selection for Discovering Distributional Treatment Effect Modifiers

Abstract

Finding the features relevant to the difference in treatment effects is essential to unveil underlying causal mechanisms. Existing methods seek such features by measuring how greatly the feature attributes affect the degree of the *conditional average treatment effect* (CATE). However, these methods may overlook important features because CATE, a measure of an average treatment effect, cannot detect the difference of other distribution parameters than the mean (e.g., variance). In this paper, we propose a feature selection framework for discovering *distributional treatment effect modifiers*. To resolve the weakness of the existing methods, we formulate a feature importance measure that quantifies how strongly the feature attributes influence the discrepancy between potential outcome distributions. We derive its computationally efficient estimator and develop a feature selection algorithm that can control the type I error rate at some desired level. Experimental results show that our framework successfully discovers important features and outperforms the existing mean-based method.

1 INTRODUCTION

When the effects of a treatment (e.g., drug administration) differ across individuals, elucidating why such heterogeneity exists is critical in many applications such as precision medicine [Lee et al., 2018], personalized education [Schochet et al., 2014], and targeted advertising [Taddy et al., 2016]. A popular approach to explaining treatment effect heterogeneity is to identify the features of an individual that are relevant to the degree of a treatment effect. For instance, to unveil the mechanism of COVID-19 vaccines, recent medical studies seek the features related to the degree of vaccine-acquired immunity [Jabal et al., 2021].

To find such features, we need to measure how greatly the attributes of each feature influence the degree of a treatment effect. To this end, the existing methods use the *conditional average treatment effect* (CATE) conditioned on each feature, i.e., an average treatment effect across the individuals who have an identical attribute of each feature [Imai and Ratkovic, 2013, Tian et al., 2014, Zhao et al., 2017]. However, this average cannot capture other distribution parameters than the mean, such as the variance. As a result, if the attributes of a feature do not affect the average treatment effect but influence other distribution parameters, these mean-based methods will incorrectly conclude that the feature is unrelated to treatment effect heterogeneity.

The goal of this paper is to propose a feature selection framework for discovering *distributional treatment effect modifiers*. To achieve this goal, we develop a feature importance measure that quantifies how greatly the attributes of each feature influence the discrepancy between the distributions of *potential outcomes*, i.e., the outcomes when an individual is treated and not treated. We formulate this measure as a variance of the maximum mean discrepancy (MMD) [Gretton et al., 2012] between the conditional potential outcome distributions conditioned on each feature. We derive its computationally efficient estimator using a kernel approximation technique and establish a feature selection algorithm that can control the type I error rate (i.e., the proportion of false-positive results) at some desired level.

Our contributions are summarized as follows:

- We formulate an MMD-based feature importance measure for discovering distributional treatment effect modifiers (Section 3.2). We derive its computationally efficient weighted estimator using a kernel approximation technique (Section 3.3).
- We develop an algorithm that selects distributional treatment effect modifiers while controlling the type I error rate (Section 3.4). To evaluate the significance, we perform multiple hypothesis tests based on the p-values computed with the conditional resampling scheme.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

- We experimentally show that our method successfully finds the features related to treatment effect heterogeneity and outperforms the existing mean-based method.

2 PRELIMINARIES

2.1 PROBLEM SETUP

Suppose that we have a sample of n individuals $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathrm{P}(A, \boldsymbol{X}, Y)$ for $i = 1, \ldots, n$. Here $A \in \{0, 1\}$ is a binary treatment ($A = 1$ if an individual is treated; otherwise, $A = 0$), $\boldsymbol{X} = [X_1, \ldots, X_d]^\top$ is d-dimensional features (a.k.a., covariates), where each feature $X_m \in \mathcal{X}$ ($m = 1, \ldots, d$) takes either discrete or continuous values, and $Y \in \mathbb{R}$ is a continuous-valued outcome. Given sample $\mathcal{D}$, we solve the problem of selecting the features in $\boldsymbol{X}$ that influence the effect of treatment A on outcome Y.

In this problem, which features should be selected depends on the measurement scale of the treatment effect [Hernán and Robins, 2020, Chapter 4]. There are two measurement scales: additive scale $Y^1 - Y^0$ and multiplicative scale Y^1 / Y^0, where Y^0 and Y^1 are random variables that are referred to as potential outcomes, each of which represents the outcome when $A = 0$ and when $A = 1$, respectively [Rubin, 1974]. In this study, we define the treatment effect for each individual on an additive scale as $Y^1 - Y^0$ because this scale is standard and widely used in numerous applications [Lee et al., 2018, Schochet et al., 2014, Taddy et al., 2016].

Unfortunately, we cannot obtain an observation of this treatment effect, $Y^1 - Y^0$. This is because we cannot jointly observe two potential outcomes Y^0 and Y^1; we only observe either Y^0 or Y^1, which is obtained as $Y = (1 - A)Y^0 + AY^1$ ($A \in \{0, 1\}$). For this reason, existing feature selection methods use the average treatment effect across individuals, which can be estimated from the data.

2.2 MEAN-BASED APPROACHES

Many existing methods seek the features whose attributes affect the degree of the average treatment effect called CATE [Tian et al., 2014, Zhao et al., 2017]. CATE is defined as a conditional expected value of the treatment effect conditioned on each feature's attribute, $X_m = x$ ($m = 1, \ldots, d$):

$$\begin{aligned} T_m(x) &:= \mathbb{E}[Y^1 - Y^0 \mid X_m = x] \\ &= \mathbb{E}[Y^1 \mid X_m = x] - \mathbb{E}[Y^0 \mid X_m = x]. \end{aligned} \tag{1}$$

CATE $T_m(x)$ is an average treatment effect over the individuals who share an identical attribute, $X_m = x$. Note that this CATE is different from the one conditioned on all features $\boldsymbol{X}$, which is an inference target of the recent causal inference methods [Johansson et al., 2016, Yoon et al., 2018].

Table 1: Joint probability tables of potential outcomes in Example 1. Nonzero probabilities are shown in bold. Total expresses marginal potential outcome probabilities.

$\mathrm{P}(Y^0, Y^1 \mid X = 0)$				
Y^0 \ Y^1	-1	0	1	Total
-1	0	0	0	0
0	**0.5**	0	**0.5**	**1.0**
1	0	0	0	0
Total	**0.5**	0	**0.5**	**1.0**

$\mathrm{P}(Y^0, Y^1 \mid X = 1)$				
Y^0 \ Y^1	-1	0	1	Total
-1	0	0	0	0
0	0	**1.0**	0	**1.0**
1	0	0	0	0
Total	0	**1.0**	0	**1.0**

Using CATE T_m ($m = 1, \ldots, d$), the features that influence the degree of the average treatment effect are defined as the following *treatment effect modifiers*:

Definition 1 (Rothman et al. [2008])**.** *Feature X_m is said to be a treatment effect modifier if there are at least two values of X_m, x_m and $x_m^\star$ ($x_m \neq x_m^\star$), such that CATE T_m in* (1) *takes different values, i.e., $T_m(x_m) \neq T_m(x_m^\star)$.*

Definition 1 states that feature X_m is a treatment effect modifier if CATE $T_m(x)$ is not a constant with respect to value $X_m = x$. Roughly speaking, when we group individuals by their X_m's values and compute the average treatment effect in each group of the individuals, if there are at least two groups with different averages, then feature X_m is a treatment effect modifier [VanderWeele, 2009].

The existing methods seek such treatment effect modifiers by fitting a regression model that is linear in treatment A with a sparse regularizer [Imai and Ratkovic, 2013, Tian et al., 2014, Zhao et al., 2017].

2.3 WEAKNESS OF MEAN-BASED APPROACHES

Since the aforementioned existing methods rely on the average treatment effect, they cannot detect the features whose attributes do not influence the average treatment effect but affect other functionals of the joint distribution of potential outcomes, such as the covariance between potential outcomes and the treatment effect variance [Russell, 2021]. To illustrate such a feature, consider the following toy example:

Example 1. *Let $Y^0, Y^1 \in \{-1, 0, 1\} \subset \mathbb{R}$ be the potential outcomes and let $X \in \{0, 1\}$ be a binary feature. Suppose that joint distribution $\mathrm{P}(Y^0, Y^1 \mid X)$ is given as Table 1. Then feature X's values are irrelevant to the average treatment effect and the covariance between potential outcomes but relevant to the treatment effect variance:*

$$\mathbb{E}[Y^1 - Y^0 \mid X = 0] = \mathbb{E}[Y^1 - Y^0 \mid X = 1] = 0$$
$$\mathrm{Cov}[Y^0, Y^1 \mid X = 0] = \mathrm{Cov}[Y^0, Y^1 \mid X = 1] = 0$$
$$\mathrm{Var}[Y^1 - Y^0 \mid X = 0] = 1; \quad \mathrm{Var}[Y^1 - Y^0 \mid X = 1] = 0.$$

Joint distribution $\mathrm{P}(Y^0, Y^1 \mid X)$ presented in Table 1 shows that feature X is related to a difference in treatment effects:

while no individual with attribute $X = 1$ receives any treatment effect, those with $X = 0$ get positive or negative effects. However, since the CATE values do not depend on X, the existing mean-based methods will incorrectly conclude that feature X is unrelated to the treatment effect heterogeneity. This implies that using CATE is insufficient to capture such *distributional* treatment effect heterogeneity and might lead to overlooking important features.

3 PROPOSED METHOD

3.1 DETECTING DISTRIBUTIONAL HETEROGENEITY

We propose a feature selection framework for discovering the features related to distributional treatment effect heterogeneity. To find such features, we consider the problem of determining whether the values of each feature X_m ($m = 1, \dots, d$) influence the functionals of the joint distribution of potential outcomes $\mathrm{P}(Y^0, Y^1 \mid X_m)$, such as the average treatment effect, the treatment effect variance, and the covariance between potential outcomes. [1] This problem is challenging because we cannot infer joint distribution $\mathrm{P}(Y^0, Y^1 \mid X_m)$ since we can never jointly observe potential outcomes Y^0 and Y^1, as described in Section 2.1.

To overcome this challenge, we propose to measure the importance of each feature X_m ($m = 1, \dots, d$) by quantifying how greatly X_m's values influence the discrepancy between conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$. This idea is motivated by the following fact: *if the discrepancy between* $\mathrm{P}(Y^0 \mid X_m)$ *and* $\mathrm{P}(Y^1 \mid X_m)$ *varies with* X_m*'s values, then joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X_m)$ *is also changeable depending on* X_m*'s values, and some functionals of the joint distribution depend on* X_m. This fact can be easily proved by taking its contraposition, as shown in Appendix A.

Such an idea enables us to detect feature X in Example 1, whose values influence the treatment effect variance. This is because, in this example, the discrepancy between conditional potential outcome distributions $\mathrm{P}(Y^0 \mid X)$ and $\mathrm{P}(Y^1 \mid X)$ changes depending on X's values.

Note, however, that our idea does not always work well. This is because there are counterexamples where feature X_m's values do not affect the discrepancy between conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ but influence joint distribution $\mathrm{P}(Y^0, Y^1 \mid X_m)$; we take a simple example in Appendix B. However, compared with the existing methods, we can detect a wider variety of features relevant to treatment effect heterogeneity, which leads to a better understanding of the underlying causal mechanisms.

[1] Identifying which functionals of the joint distribution are affected by each feature's values is extremely challenging due to the impossibility of inferring the joint distribution. One possible solution is to use the techniques for estimating the lower and upper bounds on the functionals with additional assumptions [Chen et al., 2016, Russell, 2021, Shingaki and Kuroki, 2021]. Such bounds have been successfully applied in several fields, including fairness-aware machine learning [Chikahara et al., 2021].

3.2 FEATURE IMPORTANCE MEASURE

To express the importance of each feature X_m ($m = 1, \dots, d$), we measure the discrepancy between distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ using the MMD [Gretton et al., 2012].

In fact, there are several MMD-based metrics for measuring the discrepancy between potential outcome distributions [Bellot and van der Schaar, 2021, Muandet et al., 2021, Park et al., 2021]. However, since these metrics are not designed for the conditional distributions conditioned on a single feature, they cannot be applied in our setting.

For this reason, we develop an MMD-based metric for conditional distributions $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$. Let $k_Y \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a positive-definite kernel function. Then the squared MMD between the conditional distributions conditioned on a feature value, $X_m = x$, is defined as

$$\begin{aligned} D_m^2(x) &\coloneqq \mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X_m = x), \mathrm{P}(Y^1 \mid X_m = x)) \\ &= \mathbb{E}_{Y^0, Y^{0\prime} \mid X_m = X'_m = x}[k_Y(Y^0, Y^{0\prime})] + \mathbb{E}_{Y^1, Y^{1\prime} \mid X_m = X'_m = x}[k_Y(Y^1, Y^{1\prime})] \\ &\quad - 2\,\mathbb{E}_{Y^0, Y^1 \mid X_m = x}[k_Y(Y^0, Y^1)], \end{aligned} \tag{2}$$

where superscript prime $'$ denotes an independent copy of each random variable, and expectation $\mathbb{E}_{Y^0, Y^{0\prime} \mid X_m = X'_m = x}$ is taken with respect to $\mathrm{P}(Y^0, Y^{0\prime} \mid X_m = X'_m = x)$; other expectations are taken in a similar manner. This metric has following property: if k_Y belongs to the class of kernel functions called *characteristic kernels* [Gretton et al., 2012] (e.g., the Gaussian kernel), then squared MMD is $D_m^2(x) = 0$ if and only if $\mathrm{P}(Y^0 \mid X_m = x) = \mathrm{P}(Y^1 \mid X_m = x)$.

Based on squared MMD D_m^2, we define the features related to distributional treatment effect heterogeneity as the following *distributional treatment effect modifiers*:

Definition 2. *Feature* X_m *is said to be a distributional treatment effect modifier if there are at least two values of* X_m*,* x_m *and* $x_m^\star$ *(*$x_m \neq x_m^\star$*), such that squared MMD* D_m^2 *in* (2) *takes different values, i.e.,* $D_m^2(x_m) \neq D_m^2(x_m^\star)$.

In other words, feature X_m is a distributional treatment effect modifier if the squared MMD between $\mathrm{P}(Y^0 \mid X_m)$ and $\mathrm{P}(Y^1 \mid X_m)$ varies depending on X_m's values.

To detect such a variation, we formulate the importance of each feature X_m as the variance of the squared MMD:

$$I_m \coloneqq \mathrm{Var}[D_m^2(X_m)]. \tag{3}$$

3.3 ESTIMATOR OF FEATURE IMPORTANCE

To estimate feature importance measure I_m in (3), we need to compute the expected values in (2) whose expectations

can be represented as the ones over conditional distributions $\mathrm{P}(Y^0 \mid X_m = x)$ and $\mathrm{P}(Y^1 \mid X_m = x)$.

However, we cannot directly compute them because we have no access to the observations from these conditional distributions. To overcome this difficulty, we develop a weighted estimator that can be computed from the observed data.

3.3.1 Weighted Conditional MMD (WCMMD)

To infer squared MMD $D_m^2(x)$ in (2), we develop an estimator of the expected value over conditional distribution $\mathrm{P}(Y^a \mid X_m = x)$ ($a \in \{0, 1\}$) using a weighting-based estimation technique called importance sampling.

To derive such an estimator, we use weight functions called inverse probability weights [Rosenbaum and Rubin, 1983]:

$$w_{\mathrm{P}}^0(A, \boldsymbol{X}) = \frac{\mathbf{I}(A = 0)}{1 - \mathrm{e}(\boldsymbol{X})}, \quad w_{\mathrm{P}}^1(A, \boldsymbol{X}) = \frac{\mathbf{I}(A = 1)}{\mathrm{e}(\boldsymbol{X})}, \tag{4}$$

where $\mathrm{e}(\boldsymbol{X}) := \mathrm{P}(A = 1 \mid \boldsymbol{X})$ is the conditional distribution called a *propensity score*, and $\mathbf{I}(A = a)$ is an indicator function that takes 1 if $A = a$; otherwise 0. In addition, we make the two standard assumptions: *conditional ignorability*, which requires conditional independence relation $\{Y^0, Y^1\} \perp A \mid \boldsymbol{X}$, and *positivity*, which imposes support condition $0 < \mathrm{e}(\boldsymbol{x}) < 1$ for all $\boldsymbol{x}$ [Rosenbaum and Rubin, 1983].

Under these assumptions, for instance, expected value $\mathbb{E}_{Y^1|X_m=x}[Y^1]$ can be reformulated as

$$\begin{aligned}
&\mathbb{E}_{Y^1|X_m=x}[Y^1] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m}|X_m=x}[\mathbb{E}_{Y^1|\boldsymbol{X}_{-m},X_m=x}[Y^1]] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m}|X_m=x,A=1}\left[\mathbb{E}_{Y|\boldsymbol{X}_{-m},X_m=x,A=1}\left[\frac{\mathrm{P}(A=1)}{\mathrm{P}(A=1 \mid \boldsymbol{X})}Y\right]\right] \\
&= \mathbb{E}_{A,\boldsymbol{X}_{-m},Y|X_m=x}[w_{\mathrm{P}}^1(A, \boldsymbol{X})Y],
\end{aligned}$$

where $\boldsymbol{X}_{-m} := \boldsymbol{X} \backslash X_m$ denotes the features with X_m removed.

To estimate squared MMD $D_m^2(x)$ in (2) in the same way, we formulate the following estimator, which we call a *weighted conditional MMD* (WCMMD):

$$\begin{aligned}
&\mathrm{WCMMD}^2_{X_m=x} \\
&:= \mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w_{\mathrm{P}}^0(A, \boldsymbol{X})w_{\mathrm{P}}^0(A', \boldsymbol{X}')k_Y(Y, Y')] \\
&+ \mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w_{\mathrm{P}}^1(A, \boldsymbol{X})w_{\mathrm{P}}^1(A', \boldsymbol{X}')k_Y(Y, Y')] \\
&-2\,\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=X'_m=x}[w_{\mathrm{P}}^0(A, \boldsymbol{X})w_{\mathrm{P}}^1(A', \boldsymbol{X}')k_Y(Y, Y')].
\end{aligned} \tag{5}$$

We can show that this WCMMD equals $D_m^2(x)$ under conditional ignorability and positivity assumptions:

Proposition 1. *Suppose that conditional ignorability and positivity hold. Then* $D_m^2(x) = \mathrm{WCMMD}^2_{X_m=x}$.

See Appendix C.1 for the proof. Hence, WCMMD has the same property with $D_m^2(x)$: if k_Y is a characteristic kernel, $\mathrm{WCMMD}^2_{X_m=x} = 0$ if and only if $\mathrm{P}(Y^0 \mid x) = \mathrm{P}(Y^1 \mid x)$.

3.3.2 Empirical Estimator of WCMMD

To infer $D_m^2(x)$ with estimator (5), we estimate the expected values taken over the conditional distribution conditioned on $X_m = x$ using sample $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mathrm{P}(A, \boldsymbol{X}, Y)$.

If feature X_m takes discrete values, we only have to take the averages over the individuals with $X_m = x$. Formally, by letting $\omega_i^{a,x}$ for $i = 1, \ldots, n$ and $a \in \{0, 1\}$ be

$$\omega_i^{a,x} = \frac{\mathbf{I}(x_{m,i} = x)}{\sum_{l=1}^n \mathbf{I}(x_{m,l} = x)} w_{\mathrm{P}}^a(a_i, \boldsymbol{x}_i), \tag{6}$$

we can estimate the expected values in (5) by

$$\begin{aligned}
\widehat{D}_m^2(x) := &\sum_{i=1}^n \sum_{j=1}^n \left(\omega_i^{0,x}\omega_j^{0,x} + \omega_i^{1,x}\omega_j^{1,x}\right) k_Y(y_i, y_j) \\
&- 2\sum_{i=1}^n \sum_{j=1}^n \omega_i^{0,x}\omega_j^{1,x} k_Y(y_i, y_j).
\end{aligned} \tag{7}$$

For continuous-valued feature X_m, we smoothen indicator function $\mathbf{I}$ in (6) by employing the kernel smoothing technique [Nadaraya, 1964, Watson, 1964] as follows:

$$\omega_i^{a,x} = \frac{\frac{1}{h_{X_m}} k_{X_m}(x_{m,i}, x)}{\sum_{l=1}^n \frac{1}{h_{X_m}} k_{X_m}(x_{m,l}, x)} w_{\mathrm{P}}^a(a_i, \boldsymbol{x}_i), \tag{8}$$

where the similarity between X_m's values is measured by kernel function k_{X_m} with bandwidth h_{X_m}; in our experiments, we formulate k_{X_m} as the Gaussian kernel:

$$k_{X_m}(x_m, x_m^\star) = \exp\left(-\frac{\|x_m - x_m^\star\|^2}{h_{X_m}^2}\right).$$

In both cases where $\omega_i^{a,x}$ is given as (6) and (8), we can show the consistency of estimator $\widehat{D}_m^2(x)$, i.e., convergence to the true value in the limit of infinite sample size:

Theorem 1. *Suppose that weight* $\omega_i^{a,x}$ *is given as* (6) *or* (8). *Then under the assumptions presented in Appendix C.2, we have* $\widehat{D}_m^2(x) \xrightarrow{p} D_m^2(x)$ *as* $n \to \infty$.

See Appendix C.2 for the proof. In practice, we need to estimate $\omega_i^{a,x}$ by inferring propensity score $\mathrm{e}(\boldsymbol{X}) := \mathrm{P}(A = 1 \mid \boldsymbol{X})$ with a regression model (e.g., neural network).

A drawback of estimator $\widehat{D}_m^2(x)$ in (7) is that it needs computation time $O(n^2)$ for sample size n, implying that estimating $D_m^2(x)$ for each $x = x_{m,1}, \ldots, x_{m,n}$ requires $O(n^3)$, which is impractical for large n. To resolve this issue, in what follows, we develop a computationally efficient variant of $\widehat{D}_m^2(x)$.

3.3.3 Computationally Efficient Empirical Estimator

To reduce the time of computing estimator $\widehat{D}_m^2(x)$ in (7), we employ a kernel approximation technique called random Fourier features (RFFs) [Rahimi et al., 2007].

With RFFs, we approximate kernel function $k_Y(y_i, y_j)$ in (7) as an inner product of two feature vectors:

$$k_Y(y_i, y_j) \approx \widetilde{k}_Y(y_i, y_j) = \langle z(y_i), z(y_j) \rangle_{\mathbb{R}^r}, \tag{9}$$

where $z\colon \mathbb{R} \to \mathbb{R}^r$ is a mapping that outputs a vector of the r features, where r is a hyperparameter. These r features are randomly sampled from the Fourier transform of kernel function k_Y. We formulate k_Y as a Gaussian kernel with bandwidth h_Y; in this case, feature mapping z is given as $z(y) = [\sqrt{2}\cos(\lambda_1 y + \zeta_1), \ldots, \sqrt{2}\cos(\lambda_r y + \zeta_r)]^\top$, where $\lambda_1, \ldots, \lambda_r$ are drawn from Gaussian distribution $\mathcal{N}(0, 2h_Y)$, and $\zeta_1, \ldots, \zeta_r$ are sampled from uniform distribution $\mathrm{Unif}(0, 2\pi)$, respectively [Rahimi et al., 2007].

Based on (9), we approximate estimator $\widehat{D}^2_m(x)$ in (7) as

$$\begin{aligned}\widetilde{D}^2_m(x) \coloneqq \langle \widetilde{\mu}_{Y^0|x}, \widetilde{\mu}_{Y^0|x} \rangle_{\mathbb{R}^r} + \langle \widetilde{\mu}_{Y^1|x}, \widetilde{\mu}_{Y^1|x} \rangle_{\mathbb{R}^r} \\ - 2\langle \widetilde{\mu}_{Y^0|x}, \widetilde{\mu}_{Y^1|x} \rangle_{\mathbb{R}^r}\end{aligned} \tag{10}$$

where $\widetilde{\mu}_{Y^0|x}$ and $\widetilde{\mu}_{Y^1|x}$ are the following weighted averages of the r-dimensional random feature vector:

$$\widetilde{\mu}_{Y^0|x} = \sum_{i=1}^n \omega_i^{0,x} z(y_i); \ \widetilde{\mu}_{Y^1|x} = \sum_{i=1}^n \omega_i^{1,x} z(y_i).$$

Using (10), we estimate our feature importance measure as

$$\widetilde{I}_m = \frac{1}{n-1} \sum_{\iota=1}^n \left(\widetilde{D}^2_m(x_{m,\iota}) - \frac{1}{n} \sum_{\varsigma=1}^n \widetilde{D}^2_m(x_{m,\varsigma}) \right)^2. \tag{11}$$

Computing this estimator requires $O(rn^2)$, which is feasible by setting hyperparameter r to be moderate.

3.4 FEATURE SELECTION WITH CONDITIONAL RANDOMIZATION TEST (CRT)

Using estimated importance measures $\widetilde{I}_1, \ldots, \widetilde{I}_d$, we select distributional treatment effect modifiers. To do so, we perform multiple hypothesis tests where for each $m = 1, \ldots, d$, we consider the following null and alternative hypotheses:

$$\mathcal{H}_{0,\mathrm{m}}\colon I_m = 0 \quad \text{and} \quad \mathcal{H}_{1,\mathrm{m}}\colon I_m > 0. \tag{12}$$

To decide whether to reject each null hypothesis $\mathcal{H}_{0,\mathrm{m}}$, we compute p-value p_m, i.e., the probability of obtaining test statistic I_m such that $I_m \geq \widetilde{I}_m$ under null hypothesis $\mathcal{H}_{0,\mathrm{m}}$. Evaluating this p-value requires the distribution of test statistic I_m under $\mathcal{H}_{0,\mathrm{m}}$. However, analytically deriving this distribution is extremely difficult because the asymptotic distributions of data-dependent weights $\omega_i^{0,x}$ and $\omega_i^{1,x}$ in feature importance measure $\widetilde{I}_m$ are unclear.

For this reason, we approximate the distribution of the test statistic when null hypothesis $\mathcal{H}_{0,\mathrm{m}}$ is true. To this end, we employ a synthetic dummy feature that is unrelated to treatment effect heterogeneity. To avoid changing joint distribution $\mathrm{P}(\boldsymbol{X})$, we simulate such a synthetic dummy feature corresponding to each feature X_m so that the joint distribution of the dummy feature and other observed features $\boldsymbol{X}_{-m}$ is equal to $\mathrm{P}(\boldsymbol{X})$. To achieve this, following resampling scheme called *conditional randomization test* (CRT) [Candes et al., 2018, Section F], we sample new X_m's values from the conditional distribution, $\mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ without looking at the values of treatment A and outcome Y.

Algorithm 1 Conditional Randomization Test (CRT)

Input: sample $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n$, estimated statistic $\widetilde{I}_m$
Output: p-value $\hat{p}_m$

```
Fit generative model 𝓛 to sample 𝓓.
for b = 1, ..., B do
    for i = 1, ..., n do
        Draw x_{m,i}^(b) ~ 𝓛(X_m | x_{-m,i}).
        x_i^(b) ← x_{m,i}^(b) ∪ x_{-m,i}
    end for
    Compute test statistic I_m^(b) using {(a_i, x_i^(b), y_i)}_{i=1}^n.
end for
Compute p-value p̂_m by Eq. (13).
return p̂_m
```

Algorithm 2 Proposed feature selection framework

Input: sample $\mathcal{D} = \{(a_i, \boldsymbol{x}_i, y_i)\}_{i=1}^n$, significance level α
Output: feature index set $\hat{S} \subseteq \{1, \ldots, d\}$

```
for m = 1, ..., d do
    Compute test statistic Ĩ_m with sample 𝓓.
    Compute p-value as p̂_m ← CRT(𝓓, Ĩ_m).
end for
Adjust p-values as p̂*_1, ..., p̂*_d using a multiple testing
procedure.
Select feature index set as Ŝ = {m: p̂*_m ≤ α}.
return Ŝ
```

Our CRT proceeds as illustrated in Algorithm 1. We first estimate conditional distribution $\mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ by fitting a generative model $\mathcal{L}$ to the data; in our experiments, we employ a widely-used deep generative model called the conditional variational autoencoder (CVAE) [Sohn et al., 2015]. Then for each $b = 1, \ldots, B$, we repeat the two steps: sampling n values of feature X_m as $x_{m,i}^{(b)} \sim \mathcal{L}(X_m \mid \boldsymbol{x}_{-m,i})$ ($i = 1, \ldots, n$) and using these values to compute test statistic $\widetilde{I}_m^{(b)}$. By repeating these steps, we obtain an empirical distribution of the test statistic and compute a p-value as

$$\hat{p}_m = \frac{1}{B} \sum_{b=1}^B \mathbf{I}\left(\widetilde{I}_m^{(b)} \geq \widetilde{I}_m\right). \tag{13}$$

After computing p-values $\hat{p}_1, \ldots, \hat{p}_d$, we perform multiple hypothesis tests. Since the chance of obtaining the false positives increases with the number of hypothesis tests, we control such false positives by adjusting the p-values; we used

Benjamini-Hochber (BH) adjustment procedure [Benjamini and Hochberg, 1995] in our experiments. We summarize our feature selection framework in Algorithm 2.

An advantage of applying CRT is that if the fitted generative model equals the true conditional distribution (i.e., $\mathcal{L}(X_m \mid \boldsymbol{X}_{-m}) = \mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ for all $m = 1. \ldots, d$), it can exactly control the the type I error rate to be at most significance level α [Candes et al., 2018, Section F]. Although learning such generative models is difficult, we experimentally confirmed that our method successfully controlled the type I error rate to be close to α (Section 4.2).

On the negative side, performing CRT is computationally expensive: it requires computing test statistic B times for each feature. Although this computation is embarrassingly parallelizable, it needs $O(Bdrn^2)$ in total, even with our computationally efficient estimator of the test statistic. Our future work will investigate how to further reduce the computation time; for instance, the CRT's computationally efficient variants (e.g., Liu et al. [2021]) might be helpful.

4 EXPERIMENTS

4.1 SETUP

We compared the performance of our proposed framework with the following two baselines: (1) the existing mean-based method called the selective inference method for effect modification (SI-EM) [Zhao et al., 2017] and (2) a naive variant of our method (Naive), which samples the values of a synthetic dummy feature corresponding to X_m not from conditional distribution $\mathrm{P}(X_m \mid \boldsymbol{X}_{-m})$ but from the observed X_m's values with replacement ($m = 1, \ldots, d$).

We ran all methods with significance level $\alpha = 0.05$. As regards our method and Naive, we set the number of RFFs to $r = 1000$, selected the values of kernel bandwidths $h_{X_1}, \ldots, h_{X_d}$ and h_Y using a well-known heuristic called median heuristic [Schölkopf et al., 2002], and inferred propensity score $\mathrm{e}(\boldsymbol{X})$ by fitting a feed-forward neural network that contains two linear layers with 50 neurons and Rectified Linear Unit (ReLU) activation functions. With our method, we performed a CRT with hyperparameter $B = 100$ by learning generative model $\mathcal{L}(X_m \mid \boldsymbol{X}_{-m})$ for each $m = 1, \ldots, d$. We formulated $\mathcal{L}$ as a CVAE whose encoders and decoders are given as the feed-forward neural networks that contain two linear layers with 128 neurons and ReLU functions.

4.2 SYNTHETIC DATA EXPERIMENTS

Data: We prepared synthetic datasets as follows. We drew treatment A from the Bernoulli distribution and features $\boldsymbol{X} = [X_1, \ldots, X_d]^\top$ ($d = 30$) from the Gaussian distributions:

$$A \sim \mathrm{Ber}(0.5),$$
$$\boldsymbol{X} \mid A = 0 \sim \mathcal{N}(-\mu, \boldsymbol{\Sigma}), \quad \text{and} \quad \boldsymbol{X} \mid A = 1 \sim \mathcal{N}(\mu, \boldsymbol{\Sigma}),$$

where Ber and $\mathcal{N}$ denote the Bernoulli and Gaussian distributions, respectively, $\mu = [0.2, \ldots, 0.2]^\top$ is a d-dimensional vector, and $\boldsymbol{\Sigma}$ is a $d \times d$ covariance matrix whose (i, j)-th element is $\Sigma_{i,j} = \sigma^{|i-j|}$ ($\sigma = 0.2$) for each $i, j \in \{1, \ldots, d\}$. We sampled outcome $Y = (1 - A)Y^0 + AY^1$ by generating potential outcomes Y^0 and Y^1 with the following four generation processes where five features $X_1, \ldots, X_5$ are distributional treatment effect modifiers:

- **LinMean**:

$$Y^0 \sim \mathcal{N}(-f(X_1, \ldots, X_5), 1); Y^1 \sim \mathcal{N}(f(X_1, \ldots, X_5), 1),$$

- **NonlinMean**:

$$Y^0 \sim \mathcal{N}(-g(X_1, \ldots, X_5), 1); Y^1 \sim \mathcal{N}(g(X_1, \ldots, X_5), 1),$$

- **LinVar**:

$$Y^0 \sim \mathcal{N}(-5, 1); Y^1 \sim \mathcal{N}(0, h(f(X_1, \ldots, X_5))^2),$$

- **NonlinVar**:

$$Y^0 \sim \mathcal{N}(-5, 1); Y^1 \sim \mathcal{N}(0, h(g(X_1, \ldots, X_5))^2),$$

where f, g and h are the following functions:

$$f(X_1, \ldots, X_5) = 4X_1 + 2X_2 + X_3 + 2X_4 + 4X_5,$$
$$g(X_1, \ldots, X_5) = \sum_{j=1}^{5} (X_j - 0.5)^3 + 3 \sum_{j=1}^{5} X_j - 6,$$
$$h(v) = \max(v, 1).$$

Under LinMean and NonlinMean, features $X_1, \ldots, X_5$ influence the average treatment effect whereas under LinVar and NonlinVar, $X_1, \ldots, X_5$ affect the treatment effect variance.

Results: Using these synthetic datasets, we evaluated the performance of each method. We computed a true positive rate (TPR) and a false positive rate (FPR), defined as $\frac{d_{\mathrm{TP}}}{d_{\mathrm{T}}}$ and $\frac{d_{\mathrm{FP}}}{d - d_{\mathrm{T}}}$, where $d_{\mathrm{T}} = 5$ is the number of truly relevant features, and d_{TP}, and d_{FP} are the number of truly relevant features that are correctly selected as such and the number of irrelevant features that are wrongly selected as the relevant ones, respectively. For each method, we performed 50 experiments with different synthetic datasets generated with different random numbers and computed the average and the standard deviation of TPRs and FPRs over 50 runs.

Figure 1 presents the results on the LinMean, NonlinMean, LinVar and NonlinVar datasets. With all of them, our method successfully achieved high TPRs while controlling FPRs

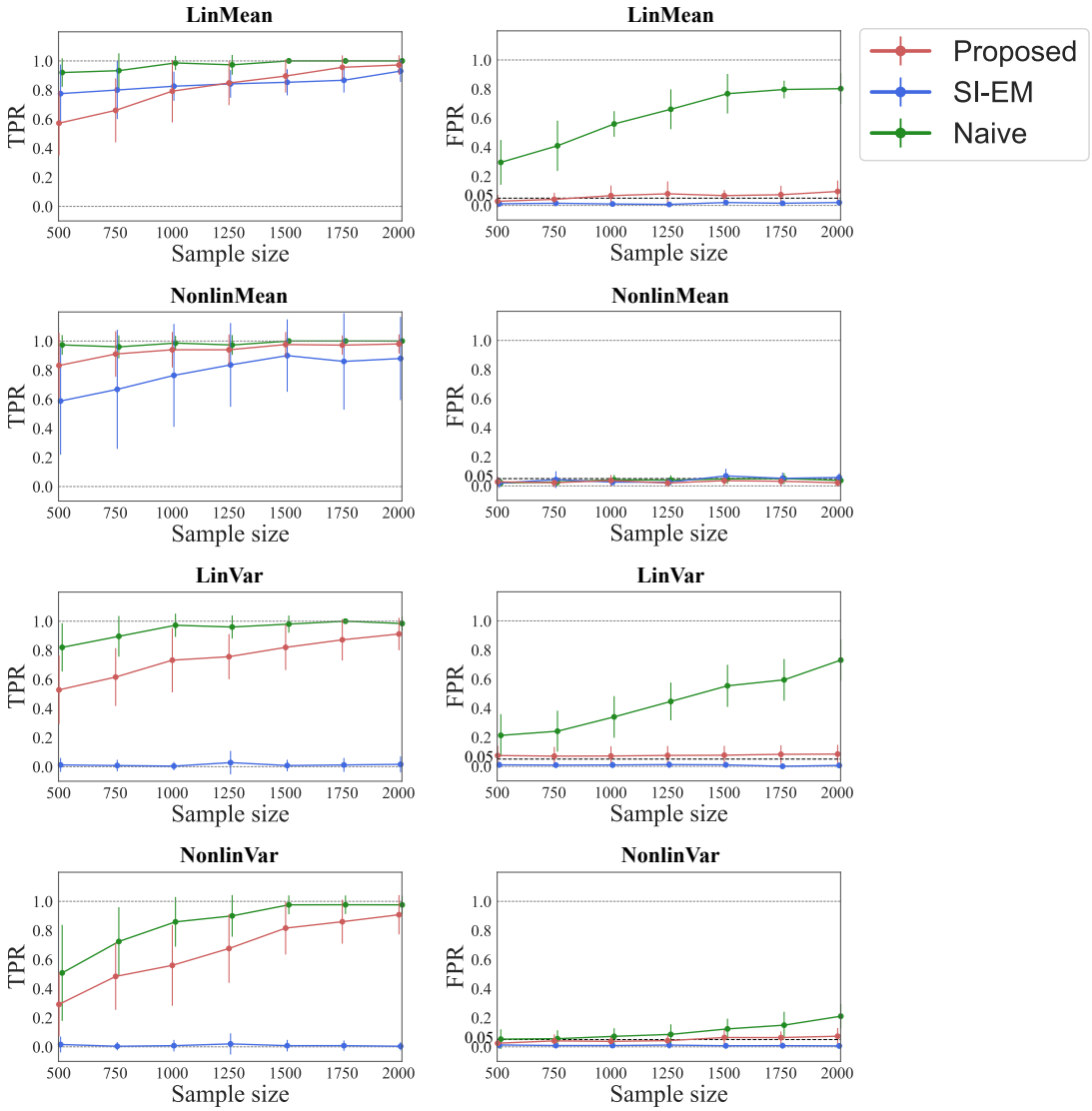

Figure 1: TPRs (left) and FPRs (right) of each method on synthetic data with sample size $n = 500, 750, \ldots, 2000$: Mean and standard deviation (error bars) over 50 runs with different datasets are shown.

to be close to $\alpha = 0.05$. Although SI-EM yielded high TPRs with the LinMean and NonlinMean datasets, since this method is not designed to detect the features related to treatment effect variance, it failed to find important features from the LinVar and NonlinVar datasets. With Naive, the FPRs were much higher than those of our method (especially with the LinMean and LinVar datasets), indicating that utilizing CRT is essential to control the FPRs.

By contrast, the use of CRT requires considerable computation time, as discussed in Section 3.4. To confirm this, we compared the run time of our method with two baselines: SI-EM and the variant of our method (Exact), which computes the feature importance measure by (7) without any approximation. Regarding our method and Exact, we evaluated the total run time, including the training time of the propensity score model and the CVAE. We ran all methods on a 64-bit CentOS machine with 2.10GHz Xeon Gold 6130 (x2) CPUs and 256-GB RAM.

Figure 2 shows the run time on the LinMean dataset with sample size $n = 500, 750, \ldots, 2000$. When $n = 2000$, SI-EM and our method required 27 and 10360 seconds, respectively, thus exhibiting a notable difference. However, our method needed far less time than Exact, demonstrating the effectiveness of kernel approximation with RFFs.

In summary, these results show the following findings:

- Our method has a computational challenge; however, it successfully discovered the features related to the average treatment effect and the treatment effect variance.
- SI-EM does not need much time; however, it failed to find the features related to the treatment effect variance.

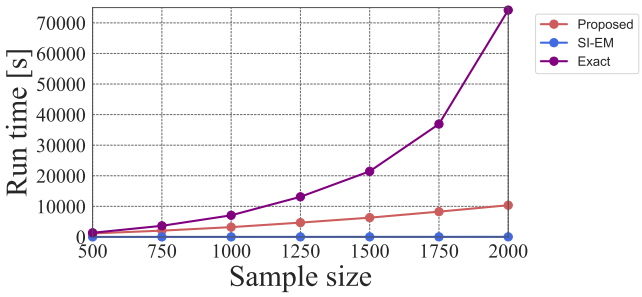

Figure 2: Run time comparison among proposed method (red), SI-EM (blue), and Exact (purple) on LinMean dataset with sample size $n = 500, 750, \ldots, 2000$

Table 2: p-values of features selected by our method from NHANES dataset: Mean and standard deviation are shown for all features with mean p-values less than $\alpha = 0.05$.

Feature	Adjusted p-value
Age	0.0075 ± 0.0305
Gender	0.0046 ± 0.0269
Amount of cigarettes smoked	0.0 ± 0.0

Thus, our proposed feature selection framework has made a significant step towards discovering the features related to distributional treatment effect heterogeneity, which, to the best of our knowledge, is the first attempt in causal inference studies. A further reduction of computation time is left as our future work, as described in Section 3.4.

4.3 REAL-WORLD DATA EXPERIMENTS

Data: We used the health records in the National Health and Nutrition Examination Survey (NHANES).[2] Following Zhao et al. [2017], we collected the records of $n = 9677$ individuals, each of which contains $d = 20$ features (3 features take continuous values, and the others are discrete).

With this dataset, we investigated which features modify the effects of obesity on low-grade systemic inflammation by regarding whether body mass index (BMI) exceeds 25 as treatment A and serum C-reactive protein (CRP) level as outcome Y. Discovering such features has important medical implications because the low-grade inflammation increases the risk of various chronic diseases, such as cancers and cardiovascular disease [Rodríguez-Hernández et al., 2013].

Since the truly relevant features are unknown, we cannot evaluate the TPRs and FPRs. For this reason, we compared the features selected by our method and SI-EM. Since our method is founded on the randomized algorithm (i.e., CRT), we computed the mean of the adjusted p-values over 50 runs and used this mean p-value to select the features.

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

Results: Table 2 presents the adjusted p-values for all features that are selected by our proposed method.

Both our method and SI-EM successfully selected age and gender, which were reported as important in the previous medical studies [Visser et al., 1999]. Although SI-EM selected only these two features, our method concluded that the amount of cigarettes smoked is also statistically significant. Selecting this feature is interesting and seems reasonable because the previous studies have reported the synergistic effect of obesity and smoking on systemic inflammation [Ólafsdóttir et al., 2005].

5 RELATED WORK

A growing number of causal inference methods have been developed to accurately estimate heterogeneous treatment effects using neural networks [Hassanpour and Greiner, 2019, Shalit et al., 2017], tree-based models [Hahn et al., 2020, Hill, 2011] and machine learning frameworks called meta-learners [Künzel et al., 2019, Nie and Wager, 2021].

However, few are designed to elucidate a causal mechanism that yields the treatment effect heterogeneity. The Causal Rule Ensemble method [Lee et al., 2020] seeks the important features by learning a rule-based model that emulates the input-output relationship of a fitted treatment effect estimation model. Gilad et al. [2021] consider a hypothesis test for discovering the treatment effect modifiers from social network data. However, none of these methods can find the features related to distributional treatment effect heterogeneity because they are also based on the average treatment effect and cannot find the features related to other functionals of the joint distribution of potential outcomes.

To overcome such a limitation of the existing mean-based methods, we established a feature selection framework for discovering the important features related to the functionals of the joint distribution of potential outcomes.

6 CONCLUSION

We proposed a feature selection framework for discovering the features related to distributional treatment effect heterogeneity. The key advantage of our framework is that it can identify the features whose values influence the functionals of the joint distribution of potential outcomes if the feature values also affect the discrepancy between conditional potential outcome distributions. To the best of our knowledge, this is the first feature selection approach to unveiling the causal mechanism that yields the distributional treatment effect heterogeneity. We experimentally show that our feature selection framework successfully selected important features and outperformed the existing method.

References

Alexis Bellot and Mihaela van der Schaar. A kernel two-sample test with selection bias. In *UAI*, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'Model-X'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80 (3):551–577, 2018.

Heng Chen, Yanqin Fan, and Ruixuan Liu. Inference for the correlation coefficient between potential outcomes in the Gaussian switching regime model. *Journal of Econometrics*, 195(2):255–270, 2016.

Yoichi Chikahara, Shinsaku Sakaue, Akinori Fujino, and Hisashi Kashima. Learning individually fair classifier with path-specific causal-effect constraint. In *AISTATS*, pages 145–153, 2021.

Amir Gilad, Harsh Parikh, Sudeepa Roy, and Babak Salimi. Heterogeneous treatment effects in social networks. *arXiv preprint arXiv:2105.10591*, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, 2012.

P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020.

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pages 5880–5887, 2019.

Miguel A. Hernán and James M. Robins. *Causal Inference: What if. Boca Raton: Chapman & Hill/CRC*. 2020.

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

Wassily Hoeffding. The strong law of large numbers for u-statistics. Technical report, 1961.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Kamal Abu Jabal, Hila Ben-Amram, Karine Beiruti, Yunis Batheesh, Christian Sussan, Salman Zarka, and Michael Edelstein. Impact of age, ethnicity, sex and prior infection status on immunogenicity following a single dose of the BNT162b2 mRNA COVID-19 vaccine: Real-world evidence from healthcare workers, Israel, December 2020 to January 2021. *Eurosurveillance*, 26(6), 2021.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, pages 3020–3029, 2016.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Kwonsang Lee, Dylan S. Small, Jesse Y. Hsu, Jeffrey H. Silber, and Paul R. Rosenbaum. Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(2): 535–546, 2018.

Kwonsang Lee, Falco J. Bargagli-Stoffi, and Francesca Dominici. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020.

Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 2021.

Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *JMLR*, 22(162):1–71, 2021.

Elizbar A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319, 2021.

Inga Sif Ólafsdóttir, Thorarinn Gislason, B. Thjodleifsson, I. Olafsson, D. Gislason, Rain Jögi, and Christer Janson. C reactive protein levels are increased in non-allergic but not allergic asthma: A multicentre epidemiological study. *Thorax*, 60(6):451–454, 2005.

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In *ICML*, pages 8401–8412, 2021.

Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NeurIPS*, volume 3, page 5, 2007.

Heriberto Rodríguez-Hernández, Luis E Simental-Mendía, Gabriela Rodríguez-Ramírez, and Miguel A Reyes-Romero. Obesity and inflammation: Epidemiology, risk

factors, and markers of inflammation. *International journal of endocrinology*, 2013.

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.

Kenneth J. Rothman, Sander Greenland, Timothy L. Lash, et al. *Modern Epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

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

Thomas M. Russell. Sharp bounds on functionals of the joint distribution in the analysis of treatment effects. *Journal of Business & Economic Statistics*, 39(2):532–546, 2021.

Peter Z. Schochet, Mike Puma, and John Deke. Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods. *National Center for Education Evaluation and Regional Assistance*, 2014.

Bernhard Schölkopf, Alexander J. Smola, Francis Bach, et al. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Robert J. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, pages 3076–3085, 2017.

Ryusei Shingaki and Manabu Kuroki. Identification and estimation of joint probabilities of potential outcomes in observational studies with covariate information. In *NeurIPS*, 2021.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NeurIPS*, pages 3483–3491, 2015.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Lu Tian, Ash A. Alizadeh, Andrew J. Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508): 1517–1532, 2014.

Tyler J. VanderWeele. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009.

Marjolein Visser, Lex M. Bouter, Geraldine M. McQuillan, Mark H. Wener, and Tamara B. Harris. Elevated C-reactive protein levels in overweight and obese adults. *Journal of the Americal Medical Association*, 282(22): 2131–2135, 1999.

Geoffrey S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.

Dominik Wied and Rafael Weißbach. Consistency of the kernel density estimator: A survey. *Statistical Papers*, 53 (1):1–21, 2012.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

Qingyuan Zhao, Dylan S. Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.

Supplementary Materials for "Feature Selection for Discovering Distributional Treatment Effect Modifiers"

A RELATIONSHIP BETWEEN MARGINAL AND JOINT DISTRIBUTIONS

To confirm that our feature importance measure is reasonable, we consider the following two relationships:

- *If the discrepancy between marginal potential outcome distributions* $\mathrm{P}(Y^0 \mid X_m)$ *and* $\mathrm{P}(Y^1 \mid X_m)$ *varies with feature* X_m*'s values, then joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X_m)$ *is also changeable depending on* X_m*'s values.*
- *If joint distribution* $\mathrm{P}(Y^0, Y^1 \mid X_m)$ *changes depending on feature* X_m*'s values, then some functionals of the joint distribution depend on* X_m*'s values.*

Since the second relationship is obvious, in this section, we show that the first relationship holds. For simplicity, we consider binary feature $X_m \in \{0, 1\}$; however, the following discussion also holds for discrete-valued and continuous-valued X_m.

To prove the first relationship, it is sufficient to show that its contraposition holds; if $\mathrm{P}(Y^0, Y^1 \mid X_m = 0) = \mathrm{P}(Y^0, Y^1 \mid X_m = 1)$, then the discrepancy between $\mathrm{P}(Y^0 \mid X_m = 0)$ and $\mathrm{P}(Y^1 \mid X_m = 0)$ equals the one between $\mathrm{P}(Y^0 \mid X_m = 1)$ and $\mathrm{P}(Y^1 \mid X_m = 1)$. We can easily prove this contraposition. From the equality of the joint distributions, we have $\mathrm{P}(Y^0 \mid X_m = 0) = \mathrm{P}(Y^0 \mid X_m = 1)$ and $\mathrm{P}(Y^1 \mid X_m = 0) = \mathrm{P}(Y^1 \mid X_m = 1)$. These equalities imply that the discrepancy between $\mathrm{P}(Y^0 \mid X_m = 0)$ and $\mathrm{P}(Y^1 \mid X_m = 0)$ equals the one between $\mathrm{P}(Y^0 \mid X_m = 1)$ and $\mathrm{P}(Y^1 \mid X_m = 1)$. Thus we proved the first relationship.

B COUNTEREXAMPLES

As described in Section 3.1, there are several counterexamples where our method cannot find the features related to the functionals of the joint distribution of potential outcomes.

Let Y^0 and Y^1 be the potential outcomes and $X \in \{0, 1\}$ be a binary feature. Suppose that the discrepancy between marginal distributions $\mathrm{P}(Y^0 \mid X)$ and $\mathrm{P}(Y^1 \mid X)$ is measured as the MMD [Gretton et al., 2012]. Then we can represent such counterexamples as the cases where the following relations hold:

$$\begin{aligned}
&\mathrm{P}(Y^0, Y^1 \mid X = 0) \neq \mathrm{P}(Y^0, Y^1 \mid X = 1) \\
&\mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 0), \mathrm{P}(Y^1 \mid X = 0)) = \mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 1), \mathrm{P}(Y^1 \mid X = 1)).
\end{aligned}$$

Letting the potential outcomes be $Y^0, Y^1 \in \{-1, 0, 1\} \subset \mathbb{R}$, we take an example of joint probability tables that satisfy the above relations in Table 3. In this example, the MMD between marginal distributions remains unchanged:

$$\mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 0), \mathrm{P}(Y^1 \mid X = 0)) = \mathrm{MMD}^2(\mathrm{P}(Y^0 \mid X = 1), \mathrm{P}(Y^1 \mid X = 1)) = 0.$$

By contrast, the joint distribution changes depending on X's values, as illustrated in Table 3. As a result, although the average treatment effect does not change, the treatment effect variance and the covariance between potential outcomes vary

Table 3: Joint probability tables of potential outcomes. Nonzero probabilities are shown in bold. Total expresses marginal potential outcome probabilities.

$P(Y^0, Y^1 \mid X = 0)$					$P(Y^0, Y^1 \mid X = 1)$				
Y^0 \ Y^1	-1	0	1	Total	Y^0 \ Y^1	-1	0	1	Total
-1	**0.5**	0	0	**0.5**	-1	0	0	**0.5**	**0.5**
0	0	0	0	0	0	0	0	0	0
1	0	0	**0.5**	**0.5**	1	**0.5**	0	0	**0.5**
Total	**0.5**	0	**0.5**	**1.0**	Total	**0.5**	0	**0.5**	**1.0**

as follows:

$$
\begin{aligned}
&\mathbb{E}[Y^1 - Y^0 \mid X = 0] = \mathbb{E}[Y^1 - Y^0 \mid X = 1] = 0 \\
&\mathrm{Cov}[Y^0, Y^1 \mid X = 0] = 1; \quad \mathrm{Cov}[Y^0, Y^1 \mid X = 1] = -1 \\
&\mathrm{Var}[Y^1 - Y^0 \mid X = 0] = 0; \quad \mathrm{Var}[Y^1 - Y^0 \mid X = 1] = 4.
\end{aligned}
$$

In this example, since we cannot detect any change in the MMD between marginal distributions, our method will fail to find that feature X is related to treatment effect heterogeneity. Note, however, that the existing mean-based approaches will also fail because the average treatment effect remains unchanged.

Addressing such counterexamples is extremely difficult. It requires us to estimate the functionals of the joint potential outcome distribution; however, inferring such a joint distribution is impossible, as described in Section 3.1. One possible solution is to utilize several techniques for estimating the lower and upper bounds on these functionals by making additional assumptions [Chen et al., 2016, Russell, 2021, Shingaki and Kuroki, 2021]. Establishing a feature selection framework that utilizes such lower and upper bounds is left as our future work.

C PROOFS

C.1 PROPOSITION 1

Proof. Recall the following definition of $\mathrm{WCMMD}^2_{X_m=x}$:

$$
\begin{aligned}
&\mathrm{WCMMD}^2_{X_m=x} \\
&:= \mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y' \mid X_m=X'_m=x}[w^0_{\mathrm{P}}(A,\boldsymbol{X})w^0_{\mathrm{P}}(A',\boldsymbol{X}')k_Y(Y,Y')] \\
&+ \mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y' \mid X_m=X'_m=x}[w^1_{\mathrm{P}}(A,\boldsymbol{X})w^1_{\mathrm{P}}(A',\boldsymbol{X}')k_Y(Y,Y')] \\
&-2\,\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y' \mid X_m=X'_m=x}[w^0_{\mathrm{P}}(A,\boldsymbol{X})w^1_{\mathrm{P}}(A',\boldsymbol{X}')k_Y(Y,Y')].
\end{aligned} \tag{5}
$$

We show that the first term in (5) equals the one in $D^2_m(x)$ in (2). Using conditional ignorability and positivity assumptions, we have

$$
\begin{aligned}
&\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y' \mid X_m=x,X'_m=x}[w^0_{\mathrm{P}}(A,\boldsymbol{X})w^0_{\mathrm{P}}(A',\boldsymbol{X}')k_Y(Y,Y')] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m} \mid X_m=x,X'_m=x}\left[\mathbb{E}_{A,A',Y,Y' \mid \boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},X_m=x,X'_m=x}\left[\frac{\mathbf{I}(A=0)}{1-\mathrm{e}(\boldsymbol{X})}\frac{\mathbf{I}(A'=0)}{1-\mathrm{e}(\boldsymbol{X}')}k_Y(Y,Y')\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m} \mid X_m=x,X'_m=x,A=0,A'=0}\left[\mathbb{E}_{Y^0,Y^{0\prime} \mid \boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},X_m=x,X'_m=x,A=0,A'=0}\left[\frac{\mathrm{P}(A=0)}{\mathrm{P}(A=0 \mid \boldsymbol{X})}\frac{\mathrm{P}(A'=0)}{\mathrm{P}(A'=0 \mid \boldsymbol{X}'))}k_Y(Y,Y')\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m} \mid X_m=x,X'_m=x}[\mathbb{E}_{Y^0,Y^{0\prime} \mid \boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},X_m=x,X'_m=x}[k_Y(Y^0,Y^{0\prime})]] \\
&= \mathbb{E}_{Y^0,Y^{0\prime} \mid X_m=x,X'_m=x}[k_Y(Y^0,Y^{0\prime})].
\end{aligned}
$$

Similarly, the second and third terms in (5) equal those in $\mathrm{MMD}^2(\mathrm{P}(Y^0 \mid x), \mathrm{P}(Y^1 \mid x))$ in (2). Thus we proved Proposition 1. □

C.2 THEOREM 1

From Proposition 1, we only have to show that $\widehat{D}_m^2(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ $(n \to \infty)$ under the assumptions of conditional ignorability and positivity:

Assumption 1 (Conditional ignorability)**.** *For treatment A, features $\boldsymbol{X}$, and potential outcomes Y^0 and Y^1, the following conditional independence relation holds:*

$$\{Y^0, Y^1\} \perp A \mid \boldsymbol{X}.$$

Assumption 2 (Positivity)**.** *For any value $\boldsymbol{x}$ of features $\boldsymbol{X}$, propensity score $\mathrm{e}(\boldsymbol{X})$ satisfies the following support condition:*

$$0 < \mathrm{e}(\boldsymbol{x}) < 1.$$

To prove $\widehat{D}_m^2(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ $(n \to \infty)$, we make several additional assumptions and impose the condition that the following symmetric function is square integrable:

$$\begin{aligned}
&K((A, \boldsymbol{X}, Y), (A', \boldsymbol{X}', Y')) \\
&:= \left(w_{\mathrm{P}}^0(A, \boldsymbol{X}) w_{\mathrm{P}}^0(A', \boldsymbol{X}') + w_{\mathrm{P}}^1(A, \boldsymbol{X}, Y) w_{\mathrm{P}}^1(A', \boldsymbol{X}', Y') - w_{\mathrm{P}}^0(A, \boldsymbol{X}) w_{\mathrm{P}}^1(A', \boldsymbol{X}') - w_{\mathrm{P}}^1(A, \boldsymbol{X}) w_{\mathrm{P}}^0(A', \boldsymbol{X}')\right) k_Y(Y, Y').
\end{aligned}$$

Assumption 3. *Symmetric function K is square integrable:*

$$\mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}[K((A, \boldsymbol{X}, Y), (A', \boldsymbol{X}', Y'))] < \infty.$$

When X_m is continuous-valued, and $\omega^{a,x}$ is given by (8), we make the following standard assumptions on kernel function k_{X_m}:

Assumption 4. *Let K_{X_m} be the following kernel function that measures the similarity between two values x_m and $x_m^\star$ on $\mathcal{X}$:*

$$K_{X_m}(x_m - x_m^\star) := \frac{1}{h_{X_m}} k_{X_m}(x_m, x_m^\star).$$

Then the order of function $K_{X_m}(u)$ is given by integer $\delta \geq 2$; in other words, the following holds:

$$\int u^\delta K_{X_m}(u) du < \infty.$$

Assumption 5. *Bandwidth h_{X_m} of kernel function k_{X_m} satisfies*

$$h_{X_m} \to 0 \quad \textit{and} \quad n h_{X_m} \to \infty. \quad (n \to \infty)$$

In addition, we impose the smoothness conditions on marginal distribution $\mathrm{P}(X_m)$ and the joint distribution of features $\mathrm{P}(\boldsymbol{X})$:

Assumption 6. *Density functions $\mathrm{P}(X_m)$ and $\mathrm{P}(\boldsymbol{X})$ are δ times continuously differentiable.*

Using these assumptions, we prove Theorem 1:

Proof. **The case where weight $\omega_i^{a,x}$ is given by** (6)**:** Let $K_{i,j} := K((a_i, \boldsymbol{x}_i, y_i), (a_j, \boldsymbol{x}_j, y_j))$ for $i, j \in \{1, \ldots, n\}$ and $n_x := \sum_{i=1}^n \mathbf{I}(x_{m,i} = x)$. Then empirical estimator $\widehat{D}_m^2(x)$ is given as

$$\begin{aligned}
\widehat{D}_m^2(x) &= \frac{1}{n_x^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{I}(x_{m,i} = x)\, \mathbf{I}(x_{m,j} = x) K_{i,j} \\
&= \left(\frac{n}{n_x}\right)^2 \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{I}(x_{m,i} = x)\, \mathbf{I}(x_{m,j} = x) K_{i,j} \\
&= \left(\frac{n}{n_x}\right)^2 V_n^x,
\end{aligned}$$

where

$$V_n^x := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{I}(x_{m,i}=x)\,\mathbf{I}(x_{m,j}=x)K_{i,j}$$

is a V-statistic whose corresponding U-statistic is given as

$$U_n^x := \frac{1}{{}_n\mathrm{C}_2}\sum_{i<j}\mathbf{I}(x_{m,i}=x)\,\mathbf{I}(x_{m,j}=x)K_{i,j}.$$

We prove the consistency of $\widehat{D}_m^2(x)$ by showing the following three relations:

$$U_n^x \xrightarrow{a.s.} \mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}[\mathbf{I}(X_m=x)\,\mathbf{I}(X_m=x)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))] \tag{14}$$

$$\left(\frac{n}{n_x}\right)^2 U_n^x \xrightarrow{a.s.} \mathrm{WCMMD}^2_{X_m=x} \tag{15}$$

$$U_n^x - V_n^x \xrightarrow{p} 0. \tag{16}$$

Relation (14) holds from the Strong Law of Large Numbers for U-statistics [Hoeffding, 1961]. By combining this relation with the fact that $\frac{n_x}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}(x_{m,i}=x) \xrightarrow{a.s.} \mathrm{P}(X_m=x)$, we can derive the relation (15). The relation in Eq. (16) can be shown as follows. Under Assumption 3, since $\mathbb{E}[K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))] \le \mathbb{E}[K((A,\boldsymbol{X},Y),(A,\boldsymbol{X},Y))] < \infty$, by employing Lemma 5.7.3 in Serfling [2009], we have $\mathbb{E}[|U_n^x - V_n^x|] = O(n^{-1})$ and hence, by applying Markov's inequality, we have

$$\mathrm{P}(|U_n^x - V_n^x| \ge \epsilon) \le \frac{\mathbb{E}[|U_n^x - V_n^x|]}{\epsilon} \to 0 \quad \text{as } n\to\infty,$$

which is sufficient to prove the relation in (16).

By combining (14), (15), and (16), we have $\widehat{D}_m^2(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ as $n\to\infty$. Since Proposition 1 holds under Assumptions 1 and 2, we have $\widehat{D}_m^2(x) \xrightarrow{p} D_m^2(x)$ as $n\to\infty$. Thus we prove the consistency of $\widehat{D}_m^2(x)$.

The case where weight $\omega_i^{a,x}$ is given by (8)**:**

In this case, empirical estimator $\widehat{D}_m^2(x)$ is given as

$$\widehat{D}_m^2(x) = \frac{\frac{1}{n^2 h_{X_m}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{X_m}(x_{m,i},x)k_{X_m}(x_{m,j},x)K_{i,j}}{\frac{1}{n^2 h_{X_m}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k_{X_m}(x_{m,i},x)k_{X_m}(x_{m,j},x)}. \tag{17}$$

From the Strong Law of Large Numbers, as $n\to\infty$, the numerator in (17) converges to the following expected value:

$$\mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}\left[\frac{1}{h_{X_m}^2}K_{X_m}\left(\frac{X_m-x}{h_{X_m}}\right)K_{X_m}\left(\frac{X_m'-x}{h_{X_m}}\right)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))\right].$$

Under Assumptions 4 and 6, we can reformulate this expected value by performing a Taylor expansion as follows:

$$\begin{aligned}
&\mathbb{E}_{A,A',\boldsymbol{X},\boldsymbol{X}',Y,Y'}\left[\frac{1}{h_{X_m}^2}K_{X_m}\left(\frac{X_m-x}{h_{X_m}}\right)K_{X_m}\left(\frac{X_m'-x}{h_{X_m}}\right)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))\right]\\
&=\mathbb{E}_{U=u,V=v}[\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=x+h_{X_m}u,X'_m=x+h_{X_m}v}[\mathrm{P}(X_m=x+h_{X_m}u)\,\mathrm{P}(X_m'=x+h_{X_m}v)K_{X_m}(u)K_{X_m}(v)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))]]\\
&=\mathbb{E}_{A,A',\boldsymbol{X}_{-m},\boldsymbol{X}'_{-m},Y,Y'|X_m=x,X'_m=x}[\mathrm{P}^2(X_m=x)K((A,\boldsymbol{X},Y),(A',\boldsymbol{X}',Y'))] + O_p\left(h_{X_m}^{\delta}\right).
\end{aligned} \tag{18}$$

Regarding the denominator in (17), from the consistency results of the kernel density estimator in Wied and Weißbach [2012], we have

$$\frac{1}{nh_{X_m}}\sum_{j=1}^{n}k_{X_m}(x_{m,j},x) \xrightarrow{a.s.} \mathrm{P}(X_m=x). \tag{19}$$

By combining (18) and (19), under Assumption 5, we have $\widehat{D}_m^2(x) \xrightarrow{p} \mathrm{WCMMD}^2_{X_m=x}$ as $n \to \infty$. Using Proposition 1, we have $\widehat{D}_m^2(x) \xrightarrow{p} D_m^2(x)$ as $n \to \infty$. Thus we proved the consistency of $\widehat{D}_m^2(x)$.

□