Feature Selection for Discovering Distributional Treatment Effect Modifiers (Supplementary material)

Yoichi Chikahara^{1,2}

Makoto Yamada²

Hisashi Kashima²

¹NTT Communication Science Laboratories, Kyoto, Japan ²Kyoto University, Kyoto, Japan

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 $P(Y^0 \mid X_m)$ and $P(Y^1 \mid X_m)$ varies with feature X_m 's values, then joint distribution $P(Y^0, Y^1 \mid X_m)$ is also changeable depending on X_m 's values.
- If joint distribution $P(Y^0, Y^1 \mid X_m)$ 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 $P(Y^0, Y^1 \mid X_m = 0) = P(Y^0, Y^1 \mid X_m = 1)$, then the discrepancy between $P(Y^0 \mid X_m = 0)$ and $P(Y^1 \mid X_m = 0)$ equals the one between $P(Y^0 \mid X_m = 1)$ and $P(Y^1 \mid X_m = 1)$. We can easily prove this contraposition. From the equality of the joint distributions, we have $P(Y^0 \mid X_m = 0) = P(Y^0 \mid X_m = 1)$ and $P(Y^1 \mid X_m = 0) = P(Y^1 \mid X_m = 1)$. These equalities imply that the discrepancy between $P(Y^0 \mid X_m = 0)$ and $P(Y^1 \mid X_m = 0)$ equals the one between $P(Y^0 \mid X_m = 1)$ and $P(Y^1 \mid X_m = 1)$. 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 $P(Y^0 \mid X)$ and $P(Y^1 \mid X)$ is measured as the MMD [Gretton et al., 2012]. Then we can represent such counterexamples as the cases where the following relations hold:

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

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 satisfies the above relations in Table A.1. In this example, the MMD between marginal distributions remains unchanged:

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

By contrast, the joint distribution changes depending on X's values, as illustrated in Table A.1. As a result, although the average treatment effect does not change, the treatment effect variance and the covariance between potential outcomes vary

Table A.1: 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	-1	0	1	Total	Y^0	-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 0.5
1	0	0	0.5	0 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:

$$\mathbb{E}[Y^1 - Y^0 \mid X = 0] = \mathbb{E}[Y^1 - Y^0 \mid X = 1] = 0$$

$$Cov[Y^0, Y^1 \mid X = 0] = 1; \quad Cov[Y^0, Y^1 \mid X = 1] = -1$$

$$Var[Y^1 - Y^0 \mid X = 0] = 0; \quad Var[Y^1 - Y^0 \mid X = 1] = 4.$$

In this example, since we cannot detect any change in the MMD between marginal distributions, our method fails to find that feature *X* is related to treatment effect heterogeneity. Note, however, that the existing mean-based approaches would 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 remains our future work.

C CHARACTERISTIC KERNELS

This section provides a brief overview on characteristic kernels. For the formal definition, see e.g., Sriperumbudur et al. [2010] and Muandet et al. [2017, Section 3.3.1].

The notion of characteristic kernels is closely related to *kernel mean embedding* [Smola et al., 2007], which is defined as the mean of feature mapping induced by a kernel function. Let $k_X \colon X \times X \to \mathbb{R}$ be a symmetric and positive-definite kernel function and $\Phi_X(x) := k_X(x,\cdot)$ be the feature mapping of kernel k_X that maps point $x \in X$ into reproducing kernel Hilbert space (RKHS) \mathcal{H}_{k_X} . Then kernel mean embedding is defined as the mean of random variable $\Phi_X(X)$:

$$\mu_X := \mathbb{E}_X[\Phi_X(X)] \in \mathcal{H}_{k_Y}.$$

Here, the expectation is taken with respect to distribution P(X); therefore, the concept of kernel mean embedding can be regarded as a mapping of distribution P(X) into the RKHS, i.e., $P(X) \mapsto \mu_X \in \mathcal{H}_{k_X}$.

A characteristic kernel is a kernel function whose kernel mean embedding does not map different distributions to the same point in the RKHS; that is, the mapping by kernel mean embedding is injective [Sriperumbudur et al., 2010].

Roughly speaking, a kernel function is characteristic if mean $\mathbb{E}_X[\Phi_X(X)]$ contains all moments of random variable X. For instance, Gaussian kernel $k_X(x, x') = \exp(-\frac{(x-x')^2}{2h_X^2})$ for $x, x' \in \mathbb{R}^1$ is characteristic because the feature mapping is given as

$$\Phi_X(x) = e^{-x^2/2h_X^2} [1, \sqrt{\frac{1}{1!h_X^2}} x, \sqrt{\frac{1}{2!h_X^4}} x^2, \dots]^\top, \text{ and its expected value } \mathbb{E}_X[\Phi_X(X)] \text{ includes all moments: } \mathbb{E}_X[X], \mathbb{E}_X[X^2], \dots$$

By contrast, if k_X is given as a polynomial function (i.e., polynomial kernel), k_X is **not** a characteristic kernel. For instance, if k_X is formulated as the 2nd-order polynomial kernel $k_X(x, x') = (1 + xx')^2$ for $x, x' \in \mathbb{R}^1$, the feature mapping is given as the finite-dimensional vector $\Phi_X(x) = [1, \sqrt{2}x, x^2]$. In this case, no element in expectation $\mathbb{E}_X[\Phi_X(X)]$ is represented as a function of higher-order moments than 2; hence, kernel k_X is not characteristic.

D PROOFS

D.1 PROPOSITION 1

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

$$WCMMD_{X_{m}=x}^{2}$$

$$\coloneqq \mathbb{E}_{A,A',X_{-m},X'_{-m},Y,Y'|X_{m}=X'_{m}=x}[w^{0}(A,X)w^{0}(A',X')k_{Y}(Y,Y')]$$

$$+ \mathbb{E}_{A,A',X_{-m},X'_{-m},Y,Y'|X_{m}=X'_{m}=x}[w^{1}(A,X)w^{1}(A',X')k_{Y}(Y,Y')]$$

$$-2 \mathbb{E}_{A,A',X_{-m},X'_{-m},Y,Y'|X_{m}=X'_{m}=x}[w^{0}(A,X)w^{1}(A',X')k_{Y}(Y,Y')]. \tag{5}$$

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

$$\begin{split} &\mathbb{E}_{A,A',X_{-m},X'_{-m},Y,Y'}|X_{m}=x,X'_{m}=x}[w^{0}(A,\boldsymbol{X})w^{0}(A',\boldsymbol{X}')k_{Y}(Y,Y')] \\ &= \mathbb{E}_{X_{-m},X'_{-m}|X_{m}=x,X'_{m}=x}\left[\mathbb{E}_{A,A',Y,Y'|X_{-m},X'_{-m},X_{m}=x,X'_{m}=x}\left[\frac{\mathbf{I}(A=0)}{1-e(\boldsymbol{X})}\frac{\mathbf{I}(A'=0)}{1-e(\boldsymbol{X}')}k_{Y}(Y,Y')\right]\right] \\ &= \mathbb{E}_{X_{-m},X'_{-m}|X_{m}=x,X'_{m}=x,A=0,A'=0}\left[\mathbb{E}_{Y^{0},Y^{0'}|X_{-m},X'_{-m},X_{m}=x,X'_{m}=x,A=0,A'=0}\left[\frac{P(A=0)}{P(A=0\mid\boldsymbol{X})}\frac{P(A'=0)}{P(A'=0\mid\boldsymbol{X}')}k_{Y}(Y,Y')\right]\right] \\ &= \mathbb{E}_{X_{-m},X'_{-m}|X_{m}=x,X'_{m}=x}[\mathbb{E}_{Y^{0},Y^{0'}|X_{-m},X'_{-m},X_{m}=x,X'_{m}=x}[k_{Y}(Y^{0},Y^{0'})]] \\ &= \mathbb{E}_{Y^{0},Y^{0'}|X_{m}=x,X'_{-m}=x}[k_{Y}(Y^{0},Y^{0'})]. \end{split}$$

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

D.2 THEOREM 1

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

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

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

Assumption 2 (Positivity). For any value x of features X, propensity score e(X) satisfies the following support condition:

$$0 < e(x) < 1$$
.

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

$$K((A, X, Y), (A', X', Y'))$$

$$:= \left(w^{0}(A, X)w^{0}(A', X') + w^{1}(A, X, Y)w^{1}(A', X', Y') - w^{0}(A, X)w^{1}(A', X') - w^{1}(A, X)w^{0}(A', X')\right)k_{Y}(Y, Y').$$

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

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

When X_m is continuous-valued, and $\omega^{a,x}$ is given by Eq. (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^* on X:

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

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

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

In addition, we impose the smoothness conditions on marginal distribution $P(X_m)$ and the joint distribution of features P(X): **Assumption 6.** Density functions $P(X_m)$ and P(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 Eq. (6): Let $K_{i,j} := K((a_i, \mathbf{x}_i, y_i), (a_j, \mathbf{x}_j, y_j))$ for $i, j \in \{1, ..., n\}$ and $n_x := \sum_{i=1}^n \mathbf{I}(x_{m,i} = x)$. Then empirical estimator $\widehat{D}_m^2(x)$ is given as

$$\widehat{D}_{m}^{2}(x) = \frac{1}{n_{x}^{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{I}(x_{m,i} = x) \mathbf{I}(x_{m,j} = x) K_{i,j}$$

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

$$= \left(\frac{n}{n_{x}}\right)^{2} V_{n}^{x},$$

where

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

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

$$U_n^x := \frac{1}{{}_{n}C_2} \sum_{i < j} \mathbf{I}(x_{m,i} = x) \mathbf{I}(x_{m,j} = x) K_{i,j}.$$

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

$$U_n^x \stackrel{a.s.}{\to} \mathbb{E}_{A,A',X,X',Y,Y'}[\mathbf{I}(X_m = x)\,\mathbf{I}(X_m = x)K((A,X,Y),(A',X',Y'))] \tag{A.1}$$

$$\left(\frac{n}{n_x}\right)^2 U_n^x \stackrel{a.s.}{\to} WCMMD_{X_m=x}^2$$
(A.2)

$$U_n^x - V_n^x \stackrel{p}{\to} 0. \tag{A.3}$$

Relation (A.1) holds from the Strong Law of Large Numbers for U-statistics [Hoeffding, 1961]. By combining this relation with the fact that $\frac{n_x}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{I}(x_{m,i} = x) \stackrel{a.s.}{\to} P(X_m = x)$, we can derive the relation in Eq. (A.2). The relation in Eq. (A.3) can be shown as follows. Under Assumption 3, since $\mathbb{E}[K((A,X,Y),(A',X',Y'))] \leq \mathbb{E}[K((A,X,Y),(A,X,Y))] < \infty$, by employing Lemma 5.7.3 in Serfling [2009], we have $\mathbb{E}[|U_n^x - V_n^x|] = O(n^{-1})$, and thus by applying Markov's inequality, we have

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

which is sufficient to prove the relation in Eq. (A.3).

By combining Eq. (A.1), (A.2), and (A.3), we have $\widehat{D}_m^2(x) \xrightarrow{p} WCMMD_{X_m=x}^2$ 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 Eq. (8):

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

$$\widehat{D}_{m}^{2}(x) = \frac{\frac{1}{n^{2}h_{\chi_{m}}^{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{X_{m}}(x_{m,i}, x) k_{\chi_{m}}(x_{m,j}, x) K_{i,j}}{\frac{1}{n^{2}h_{\chi_{m}}^{2}} \sum_{i=1}^{n} \sum_{j=1}^{n} k_{\chi_{m}}(x_{m,i}, x) k_{\chi_{m}}(x_{m,j}, x)}.$$
(A.4)

From the Strong Law of Large Numbers, as $n \to \infty$, the numerator in Eq. (A.4) converges to the following expected value:

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

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

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

$$=\mathbb{E}_{U=u,V=v}\left[\mathbb{E}_{A,A',X_{-m},X'_{-m},Y,Y'|X_{m}=x+h_{X_{m}}u,X'_{m}=x+h_{X_{m}}v}\left[P(X_{m}=x+h_{X_{m}}u)P(X'_{m}=x+h_{X_{m}}v)K_{X_{m}}(u)K_{X_{m}}(v)K((A,X,Y),(A',X',Y'))\right]\right]$$

$$=\mathbb{E}_{A,A',X_{-m},X'_{-m},X'_{-m},X'_{m}=x,X'_{m}=x}\left[P^{2}(X_{m}=x)K((A,X,Y),(A',X',Y'))\right]+O_{p}\left(h_{X_{m}}^{\delta}\right). \tag{A.5}$$

Regarding the denominator in Eq. (A.4), from the consistency results of the kernel density estimator in Wied and Weißbach [2012], we have

$$\frac{1}{nh_{X_m}} \sum_{j=1}^n k_{X_m}(x_{m,j}, x) \xrightarrow{a.s.} P(X_m = x). \tag{A.6}$$

By combining Eqs. (A.5) and (A.6), under Assumption 5, we have $\widehat{D}_m^2(x) \stackrel{p}{\to} WCMMD_{X_m=x}^2$ as $n \to \infty$. Using Proposition 1, we have $\widehat{D}_m^2(x) \stackrel{p}{\to} D_m^2(x)$ as $n \to \infty$. Thus we proved the consistency of $\widehat{D}_m^2(x)$.

E ADDITIONAL EXPERIMENTAL RESULTS

In what follows, we present several additional synthetic data experiments to further evaluate the performance of our method. Appendix E.1 shows the performance on the data where the truly relevant features do not affect the discrepancy between marginal potential outcome distributions, which is our inference target. Appendix E.2 displays the results when using different neural network architectures in the models of propensity score and CVAE.

E.1 EXAMINING COUNTEREXAMPLES

This section presents the performance of our method on the synthetic data where the features do not influence the discrepancy between conditional distributions $P(Y^0 \mid X_m)$ and $P(Y^1 \mid X_m)$ but affect joint distribution $P(Y^0, Y^1 \mid X_m)$. With such data, our method does not work well because it relies on the discrepancy between $P(Y^0 \mid X_m)$ and $P(Y^1 \mid X_m)$, as described in Section 3.1.

To evaluate the performance, we prepared synthetic data in a similar manner to Section 4.2, which only differs in the generation process of potential outcomes Y^0 and Y^1 . Here, we set the sample size to n = 2000 and sampled the values of Y^0 and Y^1 from the following 2-dimensional Gaussian distributions:

Table A.2: TPRs and FPRs of our method on LinCovar and NonlinCovar datasets. Mean and standard deviation over 50 runs are shown.

	TPR	FPR
LinCovar NonlinCovar	0.02 ± 0.06	0.02 ± 0.02
NonlinCovar	0.04 ± 0.08	0.02 ± 0.02

• LinCovar:

$$\begin{bmatrix} Y^0 \\ Y^1 \end{bmatrix} \sim \mathcal{N} \begin{pmatrix} -5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 - \frac{1}{h(f(X_1, \dots, X_5))} \\ 1 - \frac{1}{h(f(X_1, \dots, X_5))} & 1 \end{bmatrix}, \tag{A.7}$$

• NonlinCovar:

$$\begin{bmatrix} Y^0 \\ Y^1 \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} -5 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 1 - \frac{1}{h(g(X_1, \dots, X_5))} \\ 1 - \frac{1}{h(g(X_1, \dots, X_5))} & 1 \end{bmatrix} \right), \tag{A.8}$$

where functions f, g, and h are presented in Section 4.2. Under LinCovar and NonlinCovar, features X_1, \ldots, X_5 only influence the covariance between potential outcomes Y^0 and Y^1 and do not affect any functionals of the marginal distributions.

We performed 50 experiments and evaluated their mean and standard deviation of TPRs and FPRs. Table A.2 presents the results. As expected, our method could not correctly select features X_1, \ldots, X_5 because their values do not affect the discrepancy between conditional potential outcome distributions.

Note, however, that selecting these features is extremely challenging because it is impossible to estimate the covariance since we cannot infer the joint distribution of potential outcomes, as described in Section 3.1. Due to this difficulty, all of the existing mean-based methods also fail, and compared with such methods, ours can detect a wider variety of features.

E.2 PERFORMANCE EVALUATION WITH DIFFERENT NEURAL NETWORK ARCHITECTURES

Since our method relies on two neural network models to represent propensity function e(X) and CVAE $\mathcal{L}(X_m \mid X_{-m})$ (m = 1, ..., d), we confirmed how greatly the neural network architectures affect the overall feature selection performance.

For this purpose, we performed additional synthetic data experiments with sample size n = 1000. We evaluated the mean and standard deviation of TPRs and FPRs over 50 runs by changing the number of neurons of each layer in the two-layered neural network models, which is fixed to 50 for propensity score and to 128 for CVAE in the experiments in Section 4.2.

Tables A.3 and A.4 display the results. With all synthetic datasets, the number of neurons in propensity score and CVAE did not greatly affect the performance.

Table A.3: TPRs and FPRs of our method with different numbers of neurons in propensity score model. Mean and standard deviation over 50 runs are shown.

		Number of neurons in propensity score model				
		25	50	100	200	
LinMean	TPR	0.80 ± 0.21	0.79 ± 0.22	0.84 ± 0.14	0.84 ± 0.16	
	FPR	0.06 ± 0.06	0.06 ± 0.07	0.08 ± 0.06	0.08 ± 0.06	
NonlinMean	TPR	0.95 ± 0.10	0.94 ± 0.12	0.98 ± 0.06	0.97 ± 0.08	
	FPR	0.04 ± 0.04	0.04 ± 0.04	0.03 ± 0.03	0.05 ± 0.04	
LinVar	TPR	0.71 ± 0.19	0.73 ± 0.19	0.77 ± 0.16	0.76 ± 0.18	
	FPR	0.08 ± 0.07	0.07 ± 0.08	0.10 ± 0.07	0.09 ± 0.07	
NonlinVar	TPR	0.64 ± 0.25	0.62 ± 0.25	0.63 ± 0.26	0.64 ± 0.25	
	FPR	0.04 ± 0.04	0.04 ± 0.04	0.04 ± 0.04	0.04 ± 0.04	

Table A.4: TPRs and FPRs of our method with different numbers of neurons in CVAE model. Mean and standard deviation over 50 runs are shown.

		Number of neurons in CVAE model				
		16	64	128	256	
LinMean	TPR	0.82±0.18	0.82±0.17	0.79±0.22	0.83±0.16	
	FPR	0.08 ± 0.06	0.07 ± 0.06	0.06 ± 0.07	0.10 ± 0.07	
NonlinMean	TPR	0.96 ± 0.09	0.98 ± 0.06	0.94 ± 0.12	0.94 ± 0.05	
	FPR	0.04 ± 0.04	0.03 ± 0.03	0.04 ± 0.04	0.05 ± 0.04	
LinVar	TPR	0.68 ± 0.19	0.66 ± 0.17	0.73 ± 0.19	0.70 ± 0.16	
	FPR	0.07 ± 0.05	0.06 ± 0.05	0.07 ± 0.08	0.08 ± 0.07	
NonlinVar	TPR	0.58 ± 0.25	0.56 ± 0.25	0.62 ± 0.25	0.60 ± 0.20	
	FPR	0.02 ± 0.03	0.03 ± 0.03	0.04 ± 0.04	0.04 ± 0.05	

References

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.

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

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

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends*® *in Machine Learning*, 10(1-2):1–141, 2017.

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.

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

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

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

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