Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Informative Subspace Learning for Counterfactual Inference

Yale Chang, Jennifer G. Dy
Department of Electrical and Computer Engineering
Northeastern University, Boston, MA

Abstract

Inferring causal relations from observational data is widely used for knowledge discovery in healthcare and economics. To investigate whether a treatment can affect an outcome of interest, we focus on answering *counterfactual* questions of this type: what would a patient's blood pressure be had he/she recieved a different treatment? Nearest neighbor matching (NNM) sets the counterfactual outcome of any treatment (control) sample to be equal to the factual outcome of its nearest neighbor in the control (treatment) group. Although being simple, flexible and interpretable, most NNM approaches could be easily misled by variables that do not affect the outcome. In this paper, we address this challenge by learning subspaces that are predictive of the outcome variable for both the treatment group and control group. Applying NNM in the learned subspaces leads to more accurate estimation of the counterfactual outcomes and therefore treatment effects. We introduce an informative subspace learning algorithm by maximizing the nonlinear dependence between the candidate subspace and the outcome variable measured by the Hilbert-Schmidt Independence Criterion (HSIC). We propose a scalable estimator of HSIC, called *HSIC-RFF* that reduces the quadratic computational and storage complexities (with respect to the sample size) of the naive HSIC implementation to linear through constructing random Fourier features. We also prove an upper bound on the approximation error of the *HSIC-RFF* estimator. Experimental results on simulated datasets and real-world datasets demonstrate our proposed approach outperforms existing NNM approaches and other commonly used regression-based methods for counterfactual inference.

Introduction

Many challenging problems arising from healthcare, economics, and public policy depend on the discovery of reliable causal relations. For example, clinical studies aim to determine whether a new medical treatment is effective in changing clinical outcome related to a disease. Traditionally, the gold standard is to conduct randomized controlled trials (RCT) (Fisher 1960), where a set of samples are randomly assigned to the treatment group and control group. We define *treatment samples* as the set of samples in the treatment group and *control samples* as those in the control group. The treatment effect is evaluated by comparing the outcomes of treatment samples and control samples. However, RCTs are in many situations expensive or unethical. Therefore, we often need to discover causal relations from observational data.

The major challenge in causal inference from observational data is how to adjust for *confounding factors*. For example, if the ages of treatment samples are much higher than those of control samples, then the difference between outcomes of treatment and control group cannot be solely explained by the treatment. In this case, *age* is a potential confounding factor. The potential outcome framework, also called Rubin-Neyman causal model (Rubin 1974), addresses this challenge by estimating the *counterfactual outcome* of each sample. For a sample in the treatment (control) group, the counterfactual outcome is defined as the outcome when the sample is assigned to the control (treatment) group. The treatment effect can be computed given the factual outcome and estimated counterfactual outcome for each sample.

The fundamental problem in the potential outcome framework is that only the factual outcome is observed for each sample. Nearest neighbor matching (NNM) sets the counterfactual outcome of treatment (control) samples to be equal to its nearest neighbor's factual outcome in the control (treatment) group. Matching has the advantage of being simple, flexible and interpretable. However, matching heavily depends on the metric used to measure the distance between samples. Commonly used distance measures include Mahalanobis distance (Rosenbaum 2002), and propensity score (Rosenbaum and Rubin 1983; Austin 2011). It has been proved that the bias of NNM estimator increases as the dimensionality of data grows, which explains the poor performance of those matching approaches in high-dimensional data (Abadie and Imbens 2006). Li et al. proposed to reduce the bias of matching for high-dimensional data using random linear projection. However, all the existing matching approaches do not take the outcome variable into consideration, therefore could be easily misled by variables that do not affect the outcome.

To tackle the above drawback of existing NNM approaches, we propose to first learn informative subspaces that are predictive of the outcome and then apply matching in the learned subspaces. In particular, to estimate the counterfactual outcomes of treatment samples, we first learn a projection matrix by maximizing the nonlinear dependence

between the subspace and outcome variable for control samples. Then we directly apply the learned projection matrix to all the samples and find every treatment sample's matching sample in the subspace. The counterfactual outcomes of control samples can be estimated in similar procedures. Under the assumptions that 1) all the variables that affect treatment assignments and potential outcomes have been observed and included in the analysis and 2) the feature distributions of treatment group and control group have sufficient overlap, maximizing the dependence between candidate subspace and the outcome variable naturally leads to more accurate estimation of the counterfactual outcome through NNM in the learned subspace.

We use Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005; Masaeli, Dy, and Fung 2010) to measure the dependence between the candidate subspace and the outcome variable. HSIC can detect non-linear relationships due to the use of kernels. Its closed-form estimator also enables optimization w.r.t. the projection matrix. Because of the construction and storage of kernel matrices, the naive implementation of HSIC suffers from quadratic computational and storage cost w.r.t. the sample size. We propose a more efficient estimator of HSIC through approximating kernels with *random Fourier features* (Rahimi and Recht 2007). We also provide an upper bound on the approximation error of the HSIC estimator. The complexities of our new HSIC estimator, we call *HSIC-RFF*, become linear w.r.t. the sample size. Therefore, our approach can scale to large datasets.

In summary, the contributions of this work are: (1) we propose to learn subspaces that are predictive of the outcome and apply matching in the learned subspaces, therefore avoid being misled by variables that do not affect the outcome; (2) we propose a more efficient estimator of HSIC using random Fourier features, enabling our approach to be applied to large datasets; (3) experiments on simulated datasets and real-world datasets demonstrate our proposed approach outperforms existing matching methods and other commonly used regression-based methods for counterfactual inference.

In the following, we first review related work on counterfactual inference and identify the disadvantages of existing approaches. Then our proposed approach is introduced. In the experimental section, we compare our proposed approach against existing NNM methods and other regression-based methods. We summarize our work in the last section.

Related Work

The potential outcome framework is firstly proposed by (Rubin 1974). There are different methods to estimate counterfactual outcomes:

- Regression-based methods, also called covariate adjustment, explicitly model the mapping from samples' features, treatment variable to the outcome variable.
- Matching approaches match each sample to its nearest neighbors in the group that receives the opposite treatment.

Many flexible regression models (Hill 2012; Athey and Imbens 2016; Wager and Athey 2015; Johansson, Shalit, and Sontag 2016) have been used to estimate counterfactual outcomes. Compared to regression-based methods, matching approaches have the advantages of being interpretable and less sensitive to model specification (Imbens and Rubin 2015).

In this work, we will focus on nearest neighbor matching (NNM), which set the counterfactual outcome of a treatment (control) sample to be equal to the factual outcome of its nearest neighbor in the control (treatment) group, for ease of interpretation compared to its alternatives (Gu and Rosenbaum 1993). The performance of NNM heavily depends on the metric used to measure the distance between different samples. Mahalanobis distance matching (MDM) computes the Mahalanobis distance between the feature vectors of different samples and works well for low-dimensional data (Rubin 1979; Rosenbaum 2002), but it can fail when the dimensionality of features becomes large or when the distribution of features are not Gaussian (Gu and Rosenbaum 1993). Propensity score matching (PSM) (Rosenbaum and Rubin 1983) first estimates each sample's propensity score, the probability of the sample is assigned to the treatment group, using logistic regression, and then match a sample to the sample that has the closest propensity score in the group receiving the opposite treatment. PSM has been widely used for its simplicity (Dehejia and Wahba 1999; 2002; Caliendo and Kopeinig 2008). Besides logistic regression, other machine learning methods, such as neural networks (Setoguchi et al. 2008), boosting (McCaffrey, Ridgeway, and Morral 2004), and classification and regression trees (CART) (Westreich, Lessler, and Funk 2010), have also been applied to estimate the propensity scores. However, PSM was shown to be quite sensitive to the choice of variable set and sample set included in the model (Smith and Todd 2005). It can even increase the imbalance of feature distributions of treatment group and control group and therefore increase bias (King and Nielsen 2016).

In a theoretical analysis (Abadie and Imbens 2006), NNM was proved to introduce a bias term $n^{-1/d}$, where n is the number of samples and d is the number of real-valued features, for the estimation of treatment effect. Therefore, the bias of NNM approaches will increase as the number of features increases. To tackle this challenge, Li et al. proposed to apply random linear projections and run NNM in the subspaces. Johnson-Lindenstrauss (JL) lemma (Johnson and Lindenstrauss 1984) guarantees the pairwise Euclidean distances are preserved in the low-dimensional spaces.

In this work, we propose a new matching algorithm by taking the outcome variable into consideration. Note that the key difference between our approach and regression-based methods, which also use the outcome variable, is that we are not explicitly fitting any regression function. Instead, our approach has the following advantages by appying NNM in the learned subspace that is most predictive of the outcome variable (measured by HSIC): 1) compared to existing NNM methods, it avoids being misled by features that do not affect the outcome; 2) compared to regression-based methods, it is more interpretable because any sample's counterfatual outcome is directly set to be the factual outcome of its nearest neighbor in the group receiving the opposite treatment.

Proposed Method

Causal discovery aims to determine whether a treatment $\mathcal{T}$ can effectively affect the outcome of interest $\mathcal{Y}$. Given data matrix $X \in \mathbb{R}^{n\times d}$, where n is the total number of samples and d is the number of pretreatment variables, and treatment assignment vector $T \in \{0,1\}^{n\times 1}$, where $T_i = 1$ if the i-th sample receives treatment and $T_i = 0$ otherwise, the potential outcome framework (Rubin 1974) aims to estimate the individual treatment effect (ITE) for each sample

$$\mathrm{ITE}(i) = Y_i^{(1)} - Y_i^{(0)} \quad (i = 1, \cdots, n) \tag{1}$$

where $Y_i^{(1)}$ is the i-th sample's outcome if $T_i = 1$ and $Y_i^{(0)}$ is the i-th sample's outcome if $T_i = 0$. Other quantities that need to be estimated include the average treatment effect (ATE) and the average treatment effect on the treated samples (ATT) defined as following

$$\mathrm{ATE} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{ITE}(i) \tag{2}$$

$$\mathrm{ATT} = \frac{1}{n_t}\sum_{i=1}^{n}\mathbb{I}[T_i = 1]\cdot\mathrm{ITE}(i) \tag{3}$$

where n_t is the number of treated samples and $\mathbb{I}[T_i = 1]$ is equal to 1 if $T_i = 1$ and 0 otherwise. The fundamental challenge in estimating ITE is that only one outcome can be observed for each sample. We denote the observed outcomes for n samples as *factual outcomes* $Y^F \in \mathbb{R}^{n\times 1}$ and the unobserved outcomes as *counterfactual outcomes* $Y^{CF} \in \mathbb{R}^{n\times 1}$. ITE can be easily computed as follows if the counterfactual outcomes Y^{CF} can be estimated

$$\mathrm{ITE}_1^n = (Y^F - Y^{CF}) \odot (2\cdot T - 1)$$

where $\odot$ represents vector elementwise product and ITE_1^n represents the ITE values for all the n samples.

Nearest neighbor matching (NNM) is widely used to estimate counterfactual outcomes. NNM sets Y_i^{CF}, the i-th sample's counterfactual outcome, to be equal to the factual outcome of its nearest neighbor in the group receiving the opposite treatment:

$$j_{opt} = \underset{j:T_j = 1-T_i}{\operatorname{argmin}} D(x_i, x_j); \quad Y_i^{CF} = Y_{j_{opt}}^F \tag{4}$$

where x_i is the pretreatment (feature) vector of the i-th sample and $D(x_i, x_j)$ represents the distance between the i-th and j-th samples. The performance of NNM heavily depends on the choice of distance metric $D(\cdot,\cdot)$. Commonly used distance metrics include Mahanobis distance, propensity score. One key drawback of existing NNM methods is they do not take the outcome variable into consideration. As a result, they can be easily misled by variables that do not affect the outcome.

We propose to first learn subspaces that are predictive of the outcome variable and then apply NNM in the learned subspaces. In particular, for the treatment samples, we learn a projection matrix $W^{(t)} \in \mathbb{R}^{d\times q}$ to map each sample x_i from its original space $\mathbb{R}^d$ to a lower dimensional space $\mathbb{R}^q$. The embedding of x_i in the subspace can be denoted as $z_i = W^{(t)T}x_i$. *The requirement for the informative subspace is that if $||z_i - z_j||$ is small, i.e., the i-th and j-th samples are close in the subspace, their factual outcomes Y_i^F, Y_j^F should be close as well.* We achieve this requirement by maximizing the non-linear dependence between the candidate subspace $Z^{(t)} = X^{(t)}W^{(t)} \in \mathbb{R}^{n_t\times q}$ and the outcome variable $Y^{F(c)} \in \mathbb{R}^{n_t\times 1}$, where $X^{(t)} \in \mathbb{R}^{n_t\times d}$ and $Y^{F(t)}$ are the data matrix and outcomes for n_t treatment samples respectively. Many non-linear dependence measures have been proposed in the literature (Suzuki and Sugiyama 2010). We choose to use the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005) because 1) it can detect non-linear dependence relations using kernels; 2) its closed form estimator enables efficient optimization w.r.t. the projection matrix. Therefore, our objective can now be expressed as

$$\max_{W^{(t)}} \mathrm{HSIC}(Z^{(t)}, Y^{F(t)}) - \mathcal{R}(W^{(t)}) \tag{5}$$

The regularization term $\mathcal{R}(W^{(t)}) = \lambda||W||_F^2$ is used to avoid overfitting. After learning $W^{(t)}$, we can define the distance metric used in NNM as follows

$$D(x_i, x_j) = ||W^{(t)T}(x_i - x_j)||_2 \tag{6}$$

Following similar procedures, we can also learn a projection matrix $W^{(c)} \in \mathbb{R}^{d\times q}$ and estimate counterfactual outcomes for samples in the control group.

The success of our proposed approach depends on a few assumptions that are widely used in the counterfactual inference community (Imbens and Rubin 2015):

- **Unconfoundedness**: $(Y^{(1)}, Y^{(0)}) \perp\!\!\!\perp T \mid X$, the outcome variables $(Y^{(1)}, Y^{(0)})$ are independent of treatment variable T conditioned on pretreatment variables X. This assumption essentially guarantees the consistency of NNM and can be satisfied if researchers observe all the variables that affect whether a sample receives treatment and are associated with the outcomes. However, it is proved to be untestable and can be made more plausible by including more pretreatment variables in the study.
- **Overlap**: $p(T = t|X = x) > 0 \quad \forall t, x$, there should be sufficient overlap between the distributions of treatment group and control group. This assumption guarantees that we can apply the projection matrix $W^{(t)}$ learned using the treatment group to the control samples and vice versa.

Efficient HSIC Estimation

To solve for Equation (5), we need to develop an efficient estimator of HSIC. According to (Gretton et al. 2005), HSIC has a closed-form estimator defined as follows:

Definition 1 (HSIC-Gretton) *Given n independent observations $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ drawn from the joint distribution of random variables X, Y, the closed-form estimator of HSIC is defined as*

$$\mathrm{HSIC}(X, Y) := \frac{1}{n(n-1)}\mathrm{Tr}(K_X H K_Y H) \tag{7}$$

where $K_X, K_Y \in \mathbb{R}^{n\times n}$ are the kernel matrices constructed using X and Y, $H = I_n - \frac{1}{n}\mathbf{1}_{n\times n}$ is a kernel centering matrix.

However, this estimator requires the computation of kernel matrices after each update of the projection matrix, which costs $\mathcal{O}(n^2)$ in time and $\mathcal{O}(n^2)$ in space.

To address the quadratic computation complexities of the naive HSIC estimator, we propose to approximate the kernel matrices using random Fourier features (RFF) (Rahimi and Recht 2007). Given data matrix $X \in \mathbb{R}^{n\times d}$ and a shift-invariant kernel function (Scholkopf and Smola 2001) $k(\cdot,\cdot)$ that measures the similarity between samples, RFF can be constructed according to the following steps:

1. Randomly sample m parameters from a data-independent distribution $p(u)$: $u_1, \cdots, u_m \sim p(u)$, also draw m scalars from a uniform distribution: $b_1, \cdots, b_m \sim U(0, 2\pi)$;
2. Using the j-th pair (u_j, b_j), construct the j-th random feature: $f_j = [\cos(u_j^T x_1 + b_j), \cdots, \cos(u_j^T x_n + b_j)]^T$.

Note that the distribution $p(u)$ is set to be the inverse Fourier transform of the shift-invariant kernel function according to Bochner's theorem (Rudin 1962). For example, the Gaussian kernel $k(x_i, x_j) = \exp(-\frac{||x_i - x_j||_2^2}{2\sigma^2})$ can be approximated using $u_j \sim \mathcal{N}(\mathbf{0}, \frac{1}{\sigma^2}\mathbb{I}_{d\times d})$. Other commonly-used shift-invariant kernels include the Laplacian kernel, the Cauchy kernel. Note that RFF-based approximation can only be used for shift-invariant kernels. For non-shift-invariant kernels, such as the linear kernel and the polynomial kernel, we still need to use the naive HSIC estimator, which has a quadratic space and time complexity. In our experiments, we consistently use RFF to approximate Gaussian kernels and obtain encouraging results.

The random feature matrix constructed from the above steps can be written as $F = [f_1, \cdots, f_m] \in \mathbb{R}^{n\times m}$, which can be used to approximate kernel matrix (Lopez-Paz et al. 2014). We propose to approximate $\mathrm{HSIC}(X, Y)$ through RFF-based kernel approximation of K_X and K_Y. The new HSIC estimator, also called HSIC-RFF is defined as follows

Definition 2 (HSIC-RFF) *Given n independent observations $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ drawn from the joint distribution of random variables X, Y. Let the nonlinear feature matrix for X as $F \in \mathbb{R}^{n\times m}$ and that for Y as $G \in \mathbb{R}^{n\times l}$. The estimator of HSIC is defined as*

$$\begin{aligned} &\textit{HSIC-RFF}(X, Y) \qquad (8)\\ &= \frac{1}{n(n-1)}\mathrm{Tr}\left((\frac{1}{m}FF^T)H(\frac{1}{l}GG^T)H\right)\\ &= \frac{1}{mln(n-1)}||F^THG||_F^2 \end{aligned}$$

We provide an upper bound on the approximation error of the HSIC-RFF estimator using matrix Bernstein inequalities (Mackey et al. 2014).

Theorem 1 (Approximation Error Bound) *Given n independent observations $(x_1, y_1), \cdots, (x_n, y_n)$ drawn from the joint distribution of random variables X, Y and defining* HSIC *estimators* $\mathrm{HSIC}(X, Y)$, *HSIC-RFF*(X, Y) *following Definitions 1, 2, the expected difference between those two estimators has the following upper bound*

$$\mathbb{E}|\mathrm{Err}(X, Y)| \leq \frac{n}{n-1}\left(\frac{\sqrt{3n\log n}}{\sqrt{ml}} + \frac{2n\log n}{ml}\right) \qquad (9)$$

where $\mathrm{Err}(X, Y) =$ *HSIC-RFF*$(X, Y) - \mathrm{HSIC}(X, Y)$, *$n$ is the number of observations, m is the number of random features for X and l is that for Y.*

The detailed proof of this theorem is provided in the supplemental material due to space constraints. If n is fixed, the upper bound is at the scale of $\mathcal{O}(\frac{1}{\sqrt{ml}})$, which means larger value of m and l can decrease the approximation error.

Optimization

We use HSIC-RFF estimator to compute the first term in the objective Eq. (5). The final objective can be rewritten as follows:

$$\min_{W^{(t)}\in\mathbb{R}^{d\times q}} -\frac{||F^THG||_F^2}{mln_t(n_t-1)} + \lambda||W^{(t)}||_F^2 \qquad (10)$$

where $F \in \mathbb{R}^{n_t\times m}$ is the random feature matrix constructed from the treatment group's subspace $Z^{(t)} = X^{(t)}W^{(t)} \in \mathbb{R}^{n_t\times q}$, $G \in \mathbb{R}^{n_t\times l}$ is the random feature matrix constructed from the treatment group's factual outcome vector $Y^{F(t)} \in \mathbb{R}^{n_t\times 1}$, $H = I_{n_t} - \frac{1}{n_t}\mathbf{1}_{n_t\times n_t}$ is the centering matrix. Note that $W^{(c)}$, the projection matrix for the control group, can be learned in a similar manner.

Many gradient-based optimization algorithms can be used to solve this problem (Nocedal and Wright 2006). We choose a quasi-Newton algorithm, limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) because it has both super-linear convergence rate and linear memory requirement (Liu and Nocedal 1989). In order to apply L-BFGS, we need to compute the gradient of the objective w.r.t. $W^{(t)}$. For the second term, we have $\frac{\partial\lambda||W^{(t)}||_F^2}{\partial W^{(t)}} = 2\lambda W^{(t)}$. For the first term, we set $\tilde{G} = HG, A = F^T\tilde{G}$ and have the following derivations

$$\frac{\partial||F^THG||_F^2}{\partial W^{(t)}} = \sum_{i=1}^{m}\sum_{j=1}^{l} 2A_{ij}\sum_{k=1}^{n_t}\tilde{G}_{kj}\frac{\partial F_{ki}}{\partial W^{(t)}}$$

where $\frac{\partial F_{ki}}{\partial W^{(t)}} = -\sin(u_i^T W^{(t)T} x_k^{(t)} + b_i)\cdot(x_k^{(t)}u_i^T)$. Combining the gradients of the first and second term, we can compute the gradient of the objective w.r.t. the projection matrix $W^{(t)}$. $W^{(t)}$ can be initialized by sampling its entries from a normal distribution. Due to the non-convexity of our objective, we randomly initialize $W^{(t)}$ multiple times and choose the one resulting in the minimal objective value.

Complexity Analysis The cost of subspace learning is dominated by the gradient computation w.r.t. the projection matrix, which has time complexity $\mathcal{O}(n(md+ml+dq))$ and storage cost $\mathcal{O}(n(d+m+l))$. In simulations we observe that if m, l are fixed, the approximation error will not increase as sample size n increases, i.e., m, l do not need to increase as

the sample size n becomes large. Therefore, both time complexity and storage cost of subspace learning become linear w.r.t. the sample size n. In addition, all NNM approaches require nearest-neighbor searching for all the samples, which costs $\mathcal{O}(n \log n)$ using k-d tree method (Bentley 1975). Therefore, the overall time complexity of our proposed approach is $\mathcal{O}(n \cdot \kappa)$, where $\kappa = \max(\log n, md + ml + dq)$.

Experimental Results

To investigate whether our proposed approach can lead to a better distance metric for NNM, we compare our approach, also called **NNM-HSIC**, against three widely-used NNM methods:

- **MDM**: Mahalanobis distance matching (Rubin 1979);
- **PSM**: propensity score matching with logistic regression (Rosenbaum and Rubin 1983);
- **DR-RLP**: dimensionality reduction by random linear projections (Li et al. 2016).

We also compare our NNM-HSIC against three state-of-the-art regression-based approaches for counterfactual inference:

- **LASSO**: linear regression with ℓ_1 regularization to predict the factual outcome (Tibshirani 1996);
- **BART**: Bayesian additive regression trees to predict the factual outcome (Hill 2012);
- **CausalForest**: random forest regression to predict the treatment effect (Wager and Athey 2015).

Parameter Settings There are no parameters that need to be manually set for MDM and PSM. As is suggested by Li et al., we set the dimensionality of subspace to be $\lceil \log n \rceil$ and the number of random projections to be 50 for DR-RLP. For LASSO, we learn the regularization coefficient through cross-validation. For BART, we set the number of regression trees to be 200 and directly run the author's R implementation (Chipman and McCulloch 2016). For CausalForest, we also set the number of trees to be 200 and run the author's R implementation (Athey, Imbens, and Kong 2016). For our NNM-HSIC, we set the number of random features $m = l = 100$. We vary q, the dimensionality of subspace, from 1 to 10 and observe the result is not sensitive to its value in this range. Therefore we only report the result when $q = 1$. Gaussian kernel function is used to construct kernel matrices and the scale parameter σ is set using the median heuristic (Scholkopf and Smola 2001). We set the regularization coefficient $\lambda = 10^{-4}$ and the result is stable when λ is varied between 10^{-6} and 10^{-2}. We set the number of random initializations to be 20 and obtain stable results.

Evaluation Metrics Since ITE, ATE, ATT can be computed once the counterfactual outcome is estimated, we report the following three metrics to evaluate the performance of competing approaches:

- PEHE: precision in estimation of heterogeneous effect is defined as the root-mean-square error of ITE estimation, $\text{PEHE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{ITE}(x_i) - \widehat{\text{ITE}}(x_i))^2}$, where $\widehat{\text{ITE}}(x_i)$ is the estimated ITE and $\text{ITE}(x_i)$ is the true ITE for the i-th sample;
- $\mathcal{E}_{ATE}$: error in ATE estimation is defined as $\mathcal{E}_{ATE} = |\text{ATE} - \widehat{\text{ATE}}|$, where ATE is the true ATE and $\widehat{\text{ATE}}$ is the estimated ATE;
- $\mathcal{E}_{ATT}$: error in ATT estimation is defined as $\mathcal{E}_{ATT} = |\text{ATT} - \widehat{\text{ATT}}|$, where ATT is the true ATT and $\widehat{\text{ATT}}$ is the estimate ATT.

IHDP Dataset

IHDP dataset is an experimental dataset collected from the Infant Health and Development Program, a randomized experiment where intensive high-quality care were provided to low-birth-weight, premature infants. Starting from the original data, an observation study can be created by removing a nonrandom subset of the treatment group: all children with non-white mothers (Hill 2012). After preprocessing, there are 24 pretreatment variables (excluding *race*) and 747 samples in the dataset, among which there are 139 treatment samples and 608 control samples. To make the unconfoundedness assumption hold, the outcomes should be generated only using those 24 pretreatment variables and the treatment assignment. Given treatment variable $T \in \{0,1\}^{n\times 1}$ and data matrix $X \in \mathbb{R}^{n\times d}$ constructed from pretreatment variables, we generate the factual outcomes according to the following procedures suggested by Hill:

- Set $Y^{(0)} = \exp((X+C)\beta)+E_0$, where C is an offset matrix with every element equal to 0.5; $\beta \in \mathbb{R}^{d\times 1}$ is a vector of coefficients $(0, 0.1, 0.2, 0.3, 0.4)$ randomly sampled with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$; $E_0 \in \mathbb{R}^{n\times 1}$ is a vector of elements randomly sampled from the standard normal distribution $\mathcal{N}(0,1)$.
- Set $Y^{(1)} = X\beta - \omega + E_1$, where β is defined similarly to $Y^{(0)}$, $\omega \in \mathbb{R}^{n\times 1}$ is a vector with every element to some constant that makes ATT equal to 4, $E_1 \in \mathbb{R}^{n\times 1}$ is a vector of elements randomly drawn from the standard normal distribution $\mathcal{N}(0,1)$.
- Set the factual outcome vector $Y^F = Y^{(1)} \odot T + Y^{(0)} \odot (1-T)$ and the counterfactual outcome vector $Y^{CF} = Y^{(1)} \odot (1-T) + Y^{(0)} \odot T$.

Following the above procedures, the unconfoundedness assumption can be satisfied. To avoid the randomness of a single realization of the outcome variable, we repeat the above procedures 100 times and generate 100 sets of outcomes. For each outcome, we run our proposed approach and competing methods and compute PEHE, $\mathcal{E}_{ATE}$, $\mathcal{E}_{ATT}$. In this way, 100 values are generated for each evaluation metric. We report their means and standard deviations in Table 1. For each evaluation metric, we also compare the performances of different approaches using the Kruskal-Wallis test (Kruskal and Wallis 1952). Kruskal-Wallis test is a nonparametric method for testing whether two sets of samples

are drawn from the same distribution. We put the ranking of each approach, computed using Kruskal-Wallis test, in the parenthesis (smaller values indicate better performance).

	PEHE	$\mathcal{E}_{ATE}$	$\mathcal{E}_{ATT}$
MDM	7.2 ± 3.2 **(5)**	4.1 ± 0.7 **(6)**	3.2 ± 1.4 **(3)**
PSM	3.6 ± 1.4 **(2)**	0.8 ± 0.6 **(3)**	3.3 ± 1.6 **(3)**
DR-RLP	7.3 ± 3.1 **(5)**	4.1 ± 0.7 **(6)**	3.2 ± 1.4 **(3)**
LASSO	6.4 ± 3.7 **(4)**	1.6 ± 0.6 **(4)**	3.5 ± 2.1 **(3)**
BART	5.2 ± 2.9 **(3)**	1.9 ± 1.3 **(5)**	-0.1 ± 0.3 **(1)**
CausalForest	5.1 ± 2.6 **(2)**	0.4 ± 0.4 **(2)**	0.2 ± 1.0 **(2)**
NNM-HSIC	1.7 ± 0.5 **(1)**	0.1 ± 0.1 **(1)**	0.3 ± 0.3 **(2)**

Table 1: Result on *IHDP* Dataset: Each row corresponds to one approach and each column corresponds to one evaluation metric. Each entry includes mean and standard deviation of 100 metric values, followed by the ranking of the associated approach as measured by the associated evaluation metric.

As we can see, our NNM-HSIC has the best performance as measured by both PEHE and $\mathcal{E}_{ATE}$ and it is only slightly worse than BART as measured by $\mathcal{E}_{ATT}$. In particular, NNM-HSIC consistently outperforms three matching-based approaches by a large margin. This demonstrates that our proposed approach can effectively avoid being misled by variables that do not affect the outcome.

News Dataset

This dataset is firstly introduced by (Johansson, Shalit, and Sontag 2016) to simulate how the readers' experience is affected by the associated viewing device. Each sample is a piece of news represented by word counts $x_i \in \mathbb{N}^{V \times 1}$, where V is the total number of words in the vocabulary. The treatment $T_i = 1$ if the i-th sample's viewing device is mobile and $T_i = 0$ otherwise. To model the assumption that the reader prefers reading certain topics on the mobile device, a topic model can be trained on the New York Times corpus downloaded from the UCI machine learning repository (Lichman 2013). The i-th sample's outcome is assumed to be generated by $Y_i^F = C(z(x_i)^T z_0^c + T_i \cdot z(x_i)^T z_1^c) + \epsilon$, where $z(x_i)$ is the topic distribution of the i-th sample, z_0^c and z_1^c are two centroids in the topic space, $C = 50$ is a constant and ϵ is drawn from the standard normal distribution. The treatment assignment mechanism is specified by $p(T_i = 1|x_i) = \frac{\exp\left(\kappa \cdot z(x_i)^T z_1^c\right)}{\exp\left(\kappa \cdot z(x_i)^T z_0^c\right) + \exp\left(\kappa \cdot z(x_i)^T z_1^c\right)}$, where $\kappa = 10$ is a constant. We sample $n = 5000$ samples and get $d = 3477$ words. To avoid the randomness of a single realization, we repeat the generative process described above 50 times. For each realization, we run our approach and competing approaches. The results are presented in Table 2.

As we can see, our proposed NNM-HSIC has the best performance as measured by PEHE and $\mathcal{E}_{ATT}$ and it is only slightly worse than BART and LASSO measured by $\mathcal{E}_{ATE}$. Similar to the result on IHDP dataset, our approach still consistently outperforms matching-based methods by a large margin on this dataset. The only method that is comparable to NNM-HSIC on this dataset is BART. However, compared to BART, our approach is more interpretable due to the use of linear projection and nearest-neighbor matching.

	PEHE	$\mathcal{E}_{ATE}$	$\mathcal{E}_{ATT}$
MDM	4.4 ± 1.2 **(3)**	2.6 ± 0.7 **(4)**	2.6 ± 0.9 **(3)**
PSM	4.8 ± 1.2 **(3)**	2.7 ± 0.7 **(4)**	2.6 ± 1.0 **(3)**
DR-RLP	4.8 ± 1.3 **(3)**	2.7 ± 0.7 **(4)**	2.7 ± 0.9 **(3)**
LASSO	3.3 ± 1.3 **(2)**	0.2 ± 0.2 **(2)**	0.4 ± 0.4 **(1)**
BART	2.8 ± 1.0 **(1)**	0.0 ± 0.2 **(1)**	0.7 ± 0.6 **(2)**
CausalForest	4.4 ± 1.5 **(3)**	3.2 ± 1.1 **(5)**	2.5 ± 0.8 **(3)**
NNM-HSIC	2.7 ± 0.7 **(1)**	0.3 ± 0.2 **(3)**	0.5 ± 0.5 **(1)**

Table 2: Result on *News* Dataset: Each row corresponds to one approach and each column corresponds to one evaluation metric. Each entry includes mean and standard deviation of 50 metric values, followed by the ranking of the associated approach as measured by the associated evaluation metric.

Conclusion

In this work, we propose a new matching approach for counterfactual inference through learning subspaces that are predictive of the outcome variable. Compared to existing matching-based methods, our proposed approach has the advantage of not being misled by variables that do not affect the outcome. We learn the informative subspace by maximizing the nonlinear dependence between the subspace and the outcome variable measured by HSIC. To overcome the quadratic space and time complexity of the naive HSIC estimator, we propose a more efficient HSIC estimator using random Fourier features and also provide an approximation error bound. Experimental results on simulated and real-world datasets demonstrate our proposed approach outperforms existing matching-based approaches and other state-of-the-art regression-based methods for counterfactual inference.

Acknowledgements

We would like to acknowledge support for this project from the NIH grant NIH/NHLBI RO1HL089856 and RO1HL089857.

References

Abadie, A., and Imbens, G. W. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–267.

Athey, S., and Imbens, G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Athey, S.; Imbens, G.; and Kong, Y. 2016. *causalTree: Recursive Partitioning Causal Trees*. R package version 0.0.

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46(3):399–424.

Bentley, J. L. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM* 18(9):509–517.

Caliendo, M., and Kopeinig, S. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22(1):31–72.

Chipman, H., and McCulloch, R. 2016. *BayesTree: Bayesian Additive Regression Trees*. R package version 0.3-1.4.

Dehejia, R. H., and Wahba, S. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association* 94(448):1053–1062.

Dehejia, R. H., and Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84(1):151–161.

Fisher, R. A. 1960. The design of experiments.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Gu, X. S., and Rosenbaum, P. R. 1993. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* 2(4):405–420.

Hill, J. L. 2012. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*.

Imbens, G. W., and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Johansson, F. D.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. *arXiv preprint arXiv:1605.03661*.

Johnson, W. B., and Lindenstrauss, J. 1984. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics* 26(189-206):1.

King, G., and Nielsen, R. 2016. Why propensity scores should not be used for matching. 378.

Kruskal, W. H., and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47(260):583–621.

Li, S.; Vlassis, N.; Kawale, J.; and Fu, Y. 2016. Matching via dimensionality reduction for estimation of treatment effects in digital marketing campains. In *Proceedings of the 25th International Conference on Artificial Intelligence*. AAAI Press.

Lichman, M. 2013. UCI machine learning repository.

Liu, D. C., and Nocedal, J. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming* 45(1-3):503–528.

Lopez-Paz, D.; Sra, S.; Smola, A. J.; Ghahramani, Z.; and Schölkopf, B. 2014. Randomized nonlinear component analysis. In *ICML*, 1359–1367.

Mackey, L.; Jordan, M. I.; Chen, R. Y.; Farrell, B.; Tropp, J. A.; et al. 2014. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability* 42(3):906–945.

Masaeli, M.; Dy, J. G.; and Fung, G. M. 2010. From transformation-based dimensionality reduction to feature selection. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 751–758.

McCaffrey, D. F.; Ridgeway, G.; and Morral, A. R. 2004. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods* 9(4):403.

Nocedal, J., and Wright, S. 2006. *Numerical optimization*. Springer Science & Business Media.

Rahimi, A., and Recht, B. 2007. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, 1177–1184.

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

Rosenbaum, P. R. 2002. Observational studies. In *Observational Studies*. Springer. 1–17.

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

Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74(366a):318–328.

Rudin, W. 1962. Fourier analysis on groups.

Scholkopf, B., and Smola, A. J. 2001. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Setoguchi, S.; Schneeweiss, S.; Brookhart, M. A.; Glynn, R. J.; and Cook, E. F. 2008. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety* 17(6):546–555.

Smith, J. A., and Todd, P. E. 2005. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics* 125(1):305–353.

Suzuki, T., and Sugiyama, M. 2010. Sufficient dimension reduction via squared-loss mutual information estimation. In *International Conference on Artificial Intelligence and Statistics*, 804–811.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Wager, S., and Athey, S. 2015. Estimation and inference of heterogeneous treatment effects using random forests. *arXiv preprint arXiv:1510.04342*.

Westreich, D.; Lessler, J.; and Funk, M. J. 2010. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology* 63(8):826.