SCI: Subspace Learning Based Counterfactual Inference for Individual Treatment Effect Estimation

Liuyi Yao[1], Yaliang Li[1], Sheng Li[2], Mengdi Huai[3], Jing Gao[4], Aidong Zhang[3]
[1]Alibaba Group [2]University of Georgia [3]University of Virginia [4]Purdue University
{yly287738, yaliang.li}@alibaba-inc.com, sheng.li@uga.edu, {mh6ck, aidong}@virginia.edu, jinggao@purdue.edu,

ABSTRACT

Inferring causal effect from observational data has attracted much attention from various domains. Under the potential outcome framework, the estimation of counterfactuals is crucial for the investigation of causal effect at the individual level. Existing representation learning approaches focus on learning one balanced feature space, which ignores certain information predictive to the outcomes. To fully utilize the predictive information, we propose a Subspace learning based Counterfactual Inference (SCI) method to estimate causal effect at the individual level. Different from existing work, SCI learns both a common subspace, which preserves the information across all the treatment groups, and treatment-specific subspaces, which retain the information associated with each specific treatment. Learning from two kinds of subspaces helps SCI obtain better causal effect estimations than state-of-the-art methods, demonstrated by a series of experiments on synthetic and real-world datasets.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Machine learning**; • **Information systems** → **Data mining**.

KEYWORDS

Causal inference; Treatment effect estimation

ACM Reference Format:
Liuyi Yao[1], Yaliang Li[1], Sheng Li[2], Mengdi Huai[3], Jing Gao[4], Aidong Zhang[3]. 2021. SCI: Subspace Learning Based Counterfactual Inference for Individual Treatment Effect Estimation. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482175

1 INTRODUCTION

Individual Treatment Effect (ITE) is the difference between a unit's outcomes of different treatments, following the potential outcome framework [21, 23], where a unit can be any physical object, treatment is the action that applies (exposes, or subjects) to a unit, and the outcome is the result after applying the treatment. The estimated ITE facilitates decision making across various domains, such as healthcare [10], digital marketing [17], and sociology science [7]. With the estimated ITE, a doctor is able to figure out the best medication for a patient, a teacher is able to assign the most appropriate study program to a student, and a job seeker is able to decide the best training program in order to increase the employment chance. In the above examples, a *unit* could be a patient, a student, or a job seeker; The *treatments* are different medications, study programs or job training programs; The *outcomes* are the patient's status, student's test score or job seeker's employment status.

The major challenge in ITE estimation from observational data is how to handle the missing counterfactuals [1]. As a unit can only take one treatment at a time, the unit's potential outcomes of taking other treatments cannot be observed. *Counterfactuals* refer to the unobserved potential outcomes. This fact brings in the demand of estimating other treatment outcomes known as the counterfactual outcomes. However, estimating counterfactuals from the observational data encounters the issue of *treatment selection bias*. Treatment selection bias comes from the phenomenon that individuals may have different preferences on their treatment selections, resulting in distinct distributions between different treatment groups. For example, suppose people with high education degrees are prone to take the job training, the group with job training (*treated group*) would be likely to contain more highly educated people compared with the group without job training (*control group*).

There is some existing work that learns subspaces or latent representations to facilitate the estimation of counterfactuals under selection bias. Nearest neighbor matching through HSIC criteria (HSIC-NNM) [2] learns an informative subspace for each treatment and applies nearest neighbor matching on each subspace. However, HSIC-NNM only uses the corresponding treatment group to learn the subspace, which might produce unstable subspaces for data with unbalanced distributions. Other recent work focuses on learning balanced representations and inferring the outcomes based on the learned representations [13, 22]. However, the learned balanced common representations ignore some treatment-specific information which is critical for outcome prediction.

Motivated by the above challenge of learning treatment-guided and balanced representations, we propose a novel **S**ubspace learning based **C**ounterfactual **I**nference method (**SCI**). Different from existing methods, SCI learns the common subspace as well as the treatment-specific subspaces, and the two kinds of subspaces complement each other. The common subspace captures a balanced representation across different treatment groups, which preserves the information across the treatment groups and helps reduce the selection bias. Simultaneously, treatment-specific subspaces retain

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482175

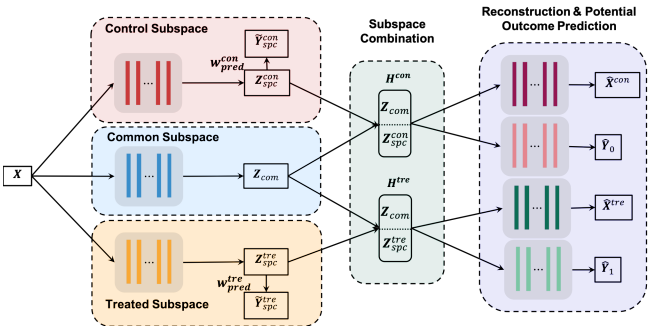

Figure 1: Framework of the proposed method SCI. The covariates X are fed into the common subspace, control subspace and treated subspace simultaneously to get the balanced common representation Z_{com} and two treatment-specific representations, Z_{spc}^{con} and Z_{spc}^{tre}. After that, SCI concatenates the common representation to every treatment-specific representation. Through the reconstruction and prediction network, the reconstructed data and two potential outcomes can be obtained.

the predictive information associated with each treatment. Consequently, the proposed method enhances the treatment outcome prediction. To demonstrate the effectiveness of the proposed method, we conduct experiments on both synthetic and real-world datasets. The experimental results confirm the benefit of learning both common and treatment-specific subspaces.

2 RELATED WORK

Existing methods for ITE estimation can be divided into two categories based on whether the ITE is estimated from one common subspace or several subspaces. The first category of methods adopts various machine learning methods to estimate ITE in one common subspace: (1) Most matching-based methods belong to this category, which map the original data to one common subspace and apply various similarity metrics to find the neighbors, such as k-nearest neighbor (k-NN) matching [5], propensity score matching [20] and deep match [14]; (2) Tree-based methods, such as Bayesian additive regression trees (BART) [5] and random forest [25], which estimate the ITE according to the leaf node that the unit belongs to; (3) Deep learning based approaches which estimate ITE based on the subspace learned by the feed-forward neural networks [12, 13, 22], variational auto-encoder [18] or generative adversarial networks [11].

The second category estimates the ITE from several subspaces. X-learner [15] adopts a meta-learning approach, which extracts the meta information (i.e., the imputed treatment effect) from treated/control subspace separately and combines the extracted information in the last procedure. HSIC-NNM [2] maps the original data to two informative subspaces related to control/treated group, respectively, and finds the nearby units in the corresponding subspace.

3 PRELIMINARY

In this work, we follow the potential outcome framework [21, 23]. First, we introduce some important notations and concepts. The dataset is denoted as $\{\mathbf{X}, \mathbf{T}, \mathbf{Y}^F\}$, where $\mathbf{X} \in \mathbb{R}^{N\times d}$ is the pre-treatment covariate matrix with N being the number of units in the dataset and d is the number of covariates. $\mathbf{T} \in \mathbb{R}^{N\times 1}$ is the treatment vector, and $\mathbf{Y}^F \in \mathbb{R}^{N\times 1}$ is the observed outcome (factual outcome). Besides, let $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$ and t_i be the pre-treatment covariate vector and the treatment of the i-th unit in the dataset, respectively. This work mainly focuses on ITE estimation under binary treatment, i.e., if $t_i = 1$, the i-th unit belongs to the treated group; otherwise, it belongs to the control group ($t_i = 0$). Before applying treatment on the i-th unit, either $y_0^{(i)}$ (control outcome) or $y_1^{(i)}$ (treated outcome), is the *potential outcome*. After applying treatment t_i, the observed outcome is the *factual outcome*, denoted as y_i^F, which equals to the potential outcome of the treatment he/she choose, i.e., $y_i^F = y_{t_i}^{(i)}$. The outcome of the other treatment is considered as the *couterfactual outcome*, denoted as y_i^{CF}, where $y_i^{CF} = y_{1-t_i}^{(i)}$.

Under the potential outcome framework, the individual treatment effect of the i-th unit is defined as: $\text{ITE}_i = y_1^{(i)} - y_0^{(i)}$, which is the difference between the potential treated and control outcomes. In this work, we develop the SCI methods under three well-known assumptions of potential outcome framework, that are stable unit treatment value assumption (SUTVA), ignorability assumption and positivity assumption [10].

4 METHODOLOGY

Overview. Estimating counterfactuals is essential in calculating the ITE. To accurately estimate the counterfactuals, we propose the **s**ubspace learning based **c**ounterfactual **i**nference method (**SCI**). Figure 1 shows the framework of SCI, which contains five major components: subspace learning including common subspace, control subspace, and treated subspace, subspace combination and the reconstruction & outcome prediction.

Objective Function. The objective function of SCI is:

$$\mathcal{L} = \mathcal{L}_f + \alpha\mathcal{L}_b + \beta\left(\mathcal{L}_{pesu}^{con} + \mathcal{L}_{pesu}^{tre}\right) + \rho\mathcal{L}_{\text{HSIC}} + \gamma\mathcal{L}_{rec} + \lambda||W||_2, \quad (1)$$

where $\mathcal{L}_f$ is the factual outcome prediction loss; $\mathcal{L}_b$ is the balancing regularization in the common subspace; $\mathcal{L}_{pesu}^{con}$ and $\mathcal{L}_{pesu}^{tre}$ are the pseudo outcome losses in the control and treated subspace, separately; $\mathcal{L}_{\text{HSIC}}$ is the dependency regularization in subspace combination; and $\mathcal{L}_{rec}$ is the reconstruction loss. The last term is the L_2 regularization on all parameters W (except the bias term). α, β, γ, ρ and λ are the hyper-parameters. The details of architecture and each term in Eq. (1) will be described in the following.

4.1 Architecture

4.1.1 Control Subspace. The goal of learning a control subspace is to investigate the treatment-specific information which is beneficial to the control outcome inference. In the control subspace, the pre-treatment feature $\mathbf{X}$ is fed into the representation network Φ_{con} to obtain the control-specific representation: $\mathbf{Z}_{spc}^{con} = \Phi_{con}(\mathbf{X})$, where Φ_{con} is a feed-forward neural network with d_{con} hidden layers and the exponential linear unit (ELU) [4] as the activation function.

To ensure that representation network $\Phi_{con}(\cdot)$ can encode control-oriented information, an external linear prediction layer is adopted to get the pseudo control outcome: $\tilde{\mathbf{Y}}_{spc}^{con} = (\mathbf{w}_{spc}^{con})'\mathbf{Z}_{spc}^{con} + b_{spc}^{con}$, where $\mathbf{Z}_{spc}^{con} \in \mathbb{R}^{k_{con}\times N}$, $\mathbf{w}_{spc}^{con} \in \mathbb{R}^{k_{con}}$, k_{con} is the dimension

of the last hidden layer of Φ_{con}. The representation $\mathbf{Z}_{spc}^{con}$ contains predictive information specifically related to control outcomes, thereby it can also be viewed as the meta-information extracted by the linear prediction procedure. Different from the existing meta-learning algorithm X-learner [15] that fixes the first step (meta-information extraction step) and estimates the ITE in its second step, we jointly optimize the control subspace learning network and the potential outcome prediction network in order to obtain better representations.

The control-specific representation and the control predictor are learned only through the control group, therefore, due to the selection bias, the predictor would not generalize well across the whole dataset. To this end, we name the predicted outcome $\tilde{\mathbf{Y}}_{spc}^{con}$ as the *pseudo control outcome*. When the pseudo control outcomes approach the factual control outcomes at a certain degree, the learned representations $\mathbf{Z}_{spc}^{con}$ can reveal the information related to the control outcome prediction. We use the pseudo-difference $\mathcal{L}_{pseu}^{con}$ to measure the distance between the pseudo control outcome and the factual control outcome, which is formulated as follows. If the outcome is continuous:

$$\mathcal{L}_{pseu}^{con} = \frac{1}{\sum_{i=1}^{N} \mathbb{I}(t_i = 0)} ||(\mathbf{Y}^F - \tilde{\mathbf{Y}}_{spc}^{con}) \cdot \mathrm{diag}(1 - \mathbf{T})||_2^2, \quad (2)$$

where $\mathbb{I}(\cdot)$ is the indicator function. If the outcome is categorical, the cross-entropy loss is adopted in Eqn. (2).

4.1.2 Treated Subspace. The treated subspace is analogical to the control subspace, and the goal of adding a treated subspace is to retain the treated-specific information which is helpful for estimating treated outcomes. The treated-representation neural network (denoted as Φ_{tre}) is built to learn the treated-specific representations: $\mathbf{Z}_{spc}^{tre} = \Phi_{tre}(\mathbf{X})$, where $\mathbf{Z}_{spc}^{tre} \in \mathbb{R}^{N \times k_{tre}}$, and k_{tre} is the dimension of the last layer of Φ_{tre}. The pseudo-difference in the treated subspace is defined as: If the outcome is continuous: $\mathcal{L}_{pseu}^{tre} = \frac{1}{\sum_{i=1}^{N} \mathbb{I}(t_i=1)} ||(\mathbf{Y}^F - \tilde{\mathbf{Y}}_{spc}^{tre}) \cdot \mathrm{diag}(\mathbf{T})||_2^2$. Similar to Eqn. (2), by minimizing the loss $\mathcal{L}_{pseu}^{tre}$, the learned representation $\mathbf{Z}_{spc}^{tre}$ is capable of preserving predictive information particularly related to treated outcome prediction.

4.1.3 Common Subspace. As mentioned in Section 4.1.1, only adopting the treatment specific representation $\mathbf{Z}_{spc}^{con}$ and $\mathbf{Z}_{spc}^{tre}$ is insufficient to obtain satisfactory outcome predictions due to generalization error brought by the existence of the selection bias. To overcome the selection bias, the control and treated subspaces should share some common information. Therefore, we introduce the common subspace as the linkage, which provides common information for control/treated subspaces. The common subspace aims to extract the cross-treatment information and reduce the selection bias. We adopt the standard feed-forward neural network with d_{com} hidden layers, $\mathbf{Z}_{com} = \Phi_{com}(\mathbf{X})$, where $\mathbf{Z}_{com} \in \mathbb{R}^{N \times k_{com}}$.

In order to reduce the selection bias, SCI adopts a balancing regularization to minimize the distribution distance between different treatment groups. In particular, we adopt the integral probability metric (IPM) [13, 19, 24] to measure the group distribution distance. The balancing regularization is then formulated as:

$$\mathcal{L}_b = \mathrm{IPM}\left(\Phi_{com}(\mathbf{X}) \cdot \mathrm{diag}(\mathbf{T}), \Phi_{com}(\mathbf{X}) \cdot \mathrm{diag}(1 - \mathbf{T})\right), \quad (3)$$

where diag() denotes the diagonal matrix, and $\Phi_{com}(\mathbf{X}) \cdot \mathrm{diag}(\mathbf{T})$, $\Phi_{com}(\mathbf{X}) \cdot \mathrm{diag}(1 - \mathbf{T})$ are the representations of the treated group and control group, respectively. By minimizing $\mathcal{L}_b$, the balanced representation can be learned in the common subspace, and the selection bias can be reduced.

4.1.4 Subspace Combination. The representations learned by the common subspace may be insufficient for outcome prediction, and the representations learned by the control/treated subspaces may be limited. To overcome the inadequacy of using a single subspace, SCI concatenates the normalized representations learned from both the common subspace and the control/treated subspaces: $\mathbf{H}^{con} = \begin{bmatrix} \mathbf{Z}_{spc}^{con} \\ \mathbf{Z}_{com} \end{bmatrix}$, $\mathbf{H}^{tre} = \begin{bmatrix} \mathbf{Z}_{spc}^{tre} \\ \mathbf{Z}_{com} \end{bmatrix}$, where $\mathbf{H}^{con}$ is the representation associated with the control outcome inference, $\mathbf{H}^{con} \in \mathbb{R}^{(k_{con}+k_{com}) \times N}$; $\mathbf{H}^{tre}$ is the representation associated with the treated outcome inference, and $\mathbf{H}^{tre} \in \mathbb{R}^{(k_{tre}+k_{com}) \times N}$. To prevent the treatment information leakage in this step, the following regularization is imposed upon the concatenated representations: $\mathcal{L}_{\mathrm{HSIC}} = \mathrm{HSIC}(\mathbf{H}^{con}, \mathbf{T}) + \mathrm{HSIC}(\mathbf{H}^{tre}, \mathbf{T})$, where HSIC denotes the Hilbert-Schmidt Independence Criterion (HSIC) [8] and $\mathbf{T}$ is the treatment assignment. By minimizing the $\mathcal{L}_{\mathrm{HSIC}}$, it forces the the final representations of $\mathbf{X}$ less dependent with the treatment assignment $\mathbf{T}$.

4.1.5 Reconstruction & Outcome Prediction. It is noticeable that the concatenated representations would be sufficient to reconstruct the original data, because it contains both cross-treatment and treatment-specific information. As a consequence, to make the concatenated representations more meaningful, SCI introduces the decoder networks Ψ_{con} and Ψ_{tre} to reconstruct the original control and treated data: $\hat{\mathbf{X}}^{con} = \Psi_{con}(\mathbf{H}^{con})$; $\hat{\mathbf{X}}^{tre} = \Psi_{tre}(\mathbf{H}^{tre})$. The reconstruction loss is calculated as follows:

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} \left((1 - t_i) ||\mathbf{X}[:, i] - \hat{\mathbf{X}}^{con}[:, i]||_2^2 + t_i ||\mathbf{X}[:, i] - \hat{\mathbf{X}}^{tre}[:, i]||_2^2 \right), \quad (4)$$

where $\mathbf{X}[:, i]$ is the i-th column of $\mathbf{X}$. Minimizing $\mathcal{L}_{rec}$ guarantees that the concatenated representations contain sufficient information about the original data $\mathbf{X}$.

We can then estimate the potential outcomes based on these concatenated representations. Let f_{con} and f_{tre} denote the predictors for control and treated outcomes, respectively. The predicted control outcome $\hat{\mathbf{Y}}_0$ and the predicted treated outcome $\hat{\mathbf{Y}}_1$ are calculated as: $\hat{\mathbf{Y}}_0 = f_{con}(\mathbf{H}^{con})$, $\hat{\mathbf{Y}}_1 = f_{tre}(\mathbf{H}^{tre})$. The factual loss for the prediction is calculated as follows. If the outcome is continuous (or if it is categorical, the cross entropy loss is adopted):

$$\begin{aligned} \mathcal{L}_f = & \frac{1}{\sum_{i=1}^{N} \mathbb{I}(t_i=0)} ||(\mathbf{Y}^F - \hat{\mathbf{Y}}_0) \cdot diag(1 - \mathbf{T})||_2^2 \\ & + \frac{1}{\sum_{i=1}^{N} \mathbb{I}(t_i=1)} ||(\mathbf{Y}^F - \hat{\mathbf{Y}}_1) \cdot diag(\mathbf{T})||_2^2. \end{aligned} \quad (5)$$

5 EXPERIMENTS

Baselines. We compare SCI with the following baselines: Least square Regression with the treatment as feature (**OLS/LR$_1$**); Separate linear regressors for each treatment group (**OLS/LR$_2$**); **HSIC-NNM**) [2]; **PSM**) [20]; **k-NN**) matching [5], Causal Forest [25], **BNN** [13], **CFR-MMD** [22], **CFR-WASS** [22], **TARNet** [22], **CE-VAE** [18]. In order to evaluate the effect of each component in

Method	Jobs ($\mathcal{R}_{\text{pol}}$)	
	Within-sample	Out-of-sample
OLS/LR$_1$	0.297 ± 0.010	0.307 ± 0.084
OLS/LR$_2$	0.295 ± 0.006	0.297 ± 0.084
HSIC-NNM	0.291 ± 0.019	0.311 ± 0.069
PSM	0.292 ± 0.019	0.307 ± 0.053
k-NN	0.230 ± 0.016	0.262 ± 0.038
Causal Forest	0.232 ± 0.018	0.224 ± 0.034
BNN	0.232 ± 0.008	0.240 ± 0.012
TARNet	0.228 ± 0.004	0.234 ± 0.012
CFR-MMD	0.213 ± 0.006	0.231 ± 0.009
CFR-WASS	0.225 ± 0.004	0.225 ± 0.010
CEVAE	0.212 ± 0.020	0.270 ± 0.045
SCI (Ours)	**0.204 ± 0.008**	**0.225 ± 0.014**
SCI-w/o-rec.	0.215 ± 0.007	0.233 ± 0.010
SCI-w/o-sub.	0.213 ± 0.006	0.231 ± 0.009
SCI-w/o-com.	0.214 ± 0.007	0.248 ± 0.010
SCI-w/o-pseu.	0.211 ± 0.006	0.237 ± 0.011
SCI-w/o-hsic.	0.208 ± 0.006	0.227 ± 0.011

Table 1: Performance comparison on Jobs Dataset.

SCI, we also compare SCI with its variants: SCI without reconstruction component (**SCI-w/o-rec.**); SCI without subspace component (**SCI-w/o-sub.**); SCI without common space component (**SCI-w/o-com.**); SCI without pseudo outcome component (**SCI-w/o-pseu.**); SCI without HSIC regularization (**SCI-w/o-hsic.**).

5.1 Experiments on Real-world Datasets

Datasets. We evaluate the proposed SCI framework on the benchmark dataset, Jobs dataset [16]. The settings of Jobs dataset is the same as the one in [6, 22].

Performance Metric. The Jobs dataset only provides the factual outcomes. The ground truth of counterfactuals and ITE are unavailable. Following the settings in [22], we adopt the policy risk to evaluate the ability of ITE estimator to support the decision making. The policy loss is defined as the loss if the units are treated according to specific policy. The smaller the policy risk is, the better the ITE estimation model can support the decision making.

Performance Analysis. Table 1 shows the performance of SCI and the baselines on 10 train/validation/test splits with 56/24/20 split ratio which is the same as [22]. From the table, it is observed that SCI outperforms baseline methods in both cases, which indicates that SCI can effectively estimate the treatment outcomes and provide better treatment decision support. Among all baselines, HISC-NNM learns subspaces that are predictive of the outcome for both control and treated groups [2], but it ignores the common information between two groups. Compared with HSIC-NNM and representation learning based baselines, SCI achieves better outcome estimations for the reason that the learned common subspace and the treatment-specific subspaces complement each other. Moreover, by comparing SCI with its variants, it can be observed that each component of SCI does contribute to the proposed model.

5.2 Experiments on Synthetic Dataset

To evaluate the robustness of SCI, we conduct experiments on a synthetic dataset with different levels of selection bias.

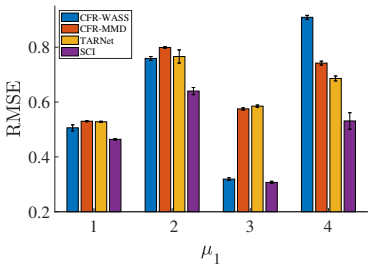

Figure 2: Performance Comparison on Synthetic Dataset.

Data Generation. Following the settings in [11], in the control group, the pre-treatment feature vector of every unit (i.e., $\mathbf{x}$) is sampled from the distribution $N(\mathbf{0}^{10\times1}, 0.5\times(\Sigma+\Sigma^T))$, where $\Sigma \sim U((-1,1)^{10\times10})$; In the treated group, every pre-treatment feature vector is generated from distribution $N(\mu_1\mathbf{1}^{10\times1}, 0.5\times(\Sigma+\Sigma^T))$. For every unit with feature $\mathbf{x}$, its control and treated outcomes are generated as follows: $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \begin{bmatrix} \mathbf{w_0}^T\mathbf{x}+n_0 \\ \mathbf{w_1}^T\mathbf{x}+n_1, \end{bmatrix}$, $\mathbf{w_0}, \mathbf{w_1} \sim U((-1,1)^{10\times1})$, and $n_0, n_1 \sim N(0, 0.1)$ where $y_0(y_1)$ is the control (treated) outcome of unit with covariate $\mathbf{x}$. Following the above procedures, by varying the value of μ_1, we can generate multiple datasets with different levels of selection bias, because the larger the μ_1 is, the smaller the overlapping of treated and control group is. Finally, we generate four datasets with $\mu_1 = 1, 2, 3, 4$, and in each dataset, there are 5,000/2,500 units in the control/treated group.

Performance Metric. On the synthetic dataset, the ground truth is known, so we adopt the precision in Estimation of Heterogeneous Effect ($\mathcal{E}_{\text{PEHE}}$) [9] as the performance metric: $\hat{\mathcal{E}}_{\text{PEHE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left([y_1^{(i)} - y_0^{(i)}] - [\hat{y}_1^{(i)} - \hat{y}_0^{(i)}]\right)^2}$, where $y_1^{(i)}$, $y_0^{(i)}$ are the ground truth outcomes, and $\hat{y}_1^{(i)}$, $\hat{y}_0^{(i)}$ are the estimated outcomes. The lower the $\hat{\mathcal{E}}_{\text{PEHE}}$ is, the better the performance.

Results Analysis. Figure 2 shows the mean and variance of RMSE on 10 realizations. The baselines we compare include TARNet, CFR-MMD and CFR-WASS, which are the most competitive baselines on the Jobs and Twins datasets. Figure 2 shows that SCI consistently outperforms its competitors under different levels of selection bias.

6 CONCLUSIONS

In this paper, we propose a novel approach for counterfactual inference by learning two kinds of subspaces with an encoder-decoder architecture. Different from existing work which learns a balanced common subspace, the proposed method SCI learns two types of subspaces: the common subspace preserves the across-treatment information and reduces the selection bias; and the treatment-specific subspaces retain the complementary information related to each treatment. Concatenating the representations learned from common and treatment-specific subspaces strengthens the ability of counterfactual inference. Extensive experiments on both synthetic and real-world datasets demonstrate the advantage of the proposed approach in counterfactual inference as well as the ITE estimation.

Acknowledgments. This work is supported in part by the US National Science Foundation under grant NSF-IIS 2008208, NSF-IIS 1955151, NSF-OAC 1934600, NSF-IIS 1938167 and NSF-IIS 1747614.

REFERENCES

[1] Ahmed Alaa and Mihaela van der Schaar. 2018. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In *Proc. of ICML'18*. 129–138.

[2] Yale Chang and Jennifer G. Dy. 2017. Informative Subspace Learning for Counterfactual Inference. In *Proc. of AAAI'17*. 1770–1776.

[3] Hugh A Chipman, Edward I George, and Robert E McCulloch. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[4] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289* (2015).

[5] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90, 3 (2008), 389–405.

[6] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84, 1 (2002), 151–161.

[7] Markus Gangl. 2010. Causal inference in sociological research. *Annual review of sociology* 36 (2010), 21–47.

[8] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

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

[10] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[11] Mihaela van der Schaar Jinsung Yoon, James Jordan. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *Proc. of ICLR'18*.

[12] Fredrik D Johansson, Nathan Kallus, Uri Shalit, and David Sontag. 2018. Learning Weighted Representations for Generalization Across Designs. *arXiv preprint arXiv:1802.08598* (2018).

[13] Fredrik D. Johansson, Uri Shalit, and David Sontag. 2016. Learning Representations for Counterfactual Inference. In *Proc. of ICML'16*.

[14] Nathan Kallus. 2018. DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training. *arXiv preprint arXiv:1802.05664* (2018).

[15] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2017. Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *arXiv preprint arXiv:1706.03461* (2017).

[16] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986), 604–620.

[17] Sheng Li, Nikos Vlassis, Jaya Kawale, and Yun Fu. 2016. Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns. In *Proc. of IJCAI'16*. 3768–3774.

[18] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal Effect Inference with Deep Latent-Variable Models. *arXiv preprint arXiv:1705.08821* (2017).

[19] Alfred Müller. 1997. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability* 29, 2 (1997), 429–443.

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

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

[22] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proc. of ICML'17*.

[23] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. 1990. On the application of probability theory to agricultural experiments. *Statist. Sci.* (1990), 465–472.

[24] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. 2012. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics* 6 (2012), 1550–1599.

[25] Stefan Wager and Susan Athey. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* just-accepted (2017).