Matching in Selective and Balanced Representation Space for Treatment Effects Estimation

by

Zhixuan Chu

(Under the Direction of Sheng Li)

Abstract

The dramatically growing availability of observational data is being witnessed in various domains of science and technology, which facilitates the study of causal inference. However, estimating treatment effects from observational data is faced with two major challenges, missing counterfactual outcomes and treatment selection bias. Matching methods are among the most widely used and fundamental approaches to estimating treatment effects, but existing matching methods have poor performance when facing data with high dimensional and complicated variables. We propose a feature selection representation matching (FSRM) method based on deep representation learning and matching, which maps the original covariate space into a selective, nonlinear, and balanced representation space, and then conducts matching in the learned representation space. We evaluate the performance of our FSRM method on three datasets, and the results demonstrate superiority over the state-of-the-art methods.

Index words: Causal Inference, Observational Data, Representation Learning, Feature Selection, Treatment Selection Bias

Matching in Selective and Balanced Representation Space for Treatment Effects Estimation

by

Zhixuan Chu

B.S., Huazhong University of Science and Technology, China, 2015

M.S., University of California, Riverside, US, 2016

A Thesis Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Science

Athens, Georgia

2021

Matching in Selective and Balanced Representation Space for Treatment Effects Estimation

by

Zhixuan Chu

Major Professor: Sheng Li

Committee: Stephen L. Rathbun

Tianming Liu

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

May 2021

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisor Dr. Li for his continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me throughout my research and the writing of this thesis. Besides, my sincere thanks goes to my committee members, Dr. Rathbun and Dr. Liu for their help on reviewing my work and giving valuable suggestions. Finally, I would like to thank my family for their constant support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Causal inference from observational data has been a critical research topic across many domains including statistics, computer science, education, public policy, economics, and health care. For example, drug developers need to know whether a new medication is beneficial or harmful in post-market surveillance; a government wants to figure out who would benefit from subsidized job training; and economists want to evaluate how a policy affects unemployment rates. Causal inference is defined as the process of estimating causal effects on units after receiving certain treatments. In the above example of drug discovery, patients are units, medication is the treatment, and the causal effect (or treatment effect) is dependent on the recovery status of patients. Although randomized controlled trials (RCT) are usually considered as the gold-standard for causal inference, estimating causal effects from observational data has become an appealing research direction owing to the increasing availability of data and the low costs.

Researchers have developed various frameworks to causal inference, and the most representative ones include the potential outcome framework (Rubin, 1974; Splawa-Neyman et al., 1990) and the structural causal model (Pearl, 1995, 2009, 2014). The potential outcome framework aims to estimate all potential outcomes under different treatments and then

calculate the treatment effects. The structural causal model (SCM) combines components of structural equations models, graphical models, and the potential outcomes framework to make causal inference. In this thesis, we focus on the potential outcome framework.

When estimating treatment effects from observational data, we face two major challenges (Yao et al., 2020), i.e., missing counterfactual outcomes and treatment selection bias. Firstly, in real life, we only observe the factual outcome and never all potential outcomes that would potentially have happened had we chosen other different treatment options. In medicine, for example, we only observe the outcome of giving a patient a specific treatment, but we never observe what would have happened if the patient was instead given an alternative treatment. Secondly, unlike randomized controlled experiments, treatments are typically not assigned at random in observational data. In the medical setting, physicians take a set of factors into account, such as the patient's feedback to the treatment, medical history, and patient health condition, when choosing a therapeutic option. Due to this treatment assignment bias, the treated population may differ significantly from the general population. These two major issues make treatment effects estimation very challenging.

A widely used solution is matching method, where the missing counterfactual outcome of a unit to a treatment is estimated by the factual outcome of its most similar neighbors that have received that treatment. The dataset including matched samples mimics a randomized controlled trial where the distribution of covariates will be similar between treatment and control groups. The only expected difference between the treatment and control groups is the outcome variable being studied. Compared to regression-based methods such as counterfactual regression (Shalit et al., 2017) and Bayesian additive regression trees (Chipman et al., 2010), matching approaches are more interpretable and less sensitive to model specification (Imbens & Rubin, 2015).

Most of existing matching methods are performed in the original covariate space (e.g., Nearest Neighbor Matching (Rubin, 1973), Coarsened Exact Matching (Iacus et al., 2012))

or in the one-dimensional propensity score space (e.g., Propensity Score Matching (Rosenbaum & Rubin, 1983)). Although rich information is retained in the original covariate space, it will face the curse of dimensionality and introduce more bias when controlling for irrelevant variables. Theoretical studies revealed that the bias of matching methods increases with the dimensionality of the covariate space (Abadie & Imbens, 2006). Propensity score matching combats the curse of dimensionality of matching directly on the original covariates by matching on the probability of a unit being assigned to a particular treatment given a set of observed covariates. However, a one-dimensional propensity score space will lose most of the information in the data. In addition, provided that models are not over-specified, nonlinear models are usually more capable of dealing with complicated data distributions.

Therefore, learning a low-dimensional balanced and nonlinear representations instead of high-dimensional original covariates space or one-dimensional propensity score space for observational data is a promising solution, which has been discussed in (Chang & Dy, 2017; S. Li & Fu, 2017). The major drawback of existing causal inference methods including matching methods is that they always treat all observed variables as pre-treatment variables, which are not affected by treatment assignments but may be predictive of outcomes. This assumption is not tenable for observational data such as post-market pharmaceutical surveillance, cross sectional studies, and electronic medical records and so on. If all observed variables are directly used to estimate treatment effects, more bias may be introduced into the model. For example, conditioning on an instrumental variable, which is associated with the treatment assignment but not with the outcome except through exposure, can increase both bias and variance of estimated treatment effects (Myers et al., 2011). cent Therefore, conditioning on these variables, let alone irrelevant variables, will introduce more impalpable bias into model, especially in scenarios with high dimensional variables.

To address the above issues, we propose a deep feature selection representation matching (FSRM) model for treatment effects estimation in a representation space. The key idea

of FSRM is to map the original covariate space into the selective, nonlinear, and balanced representation space, which is predictive of treatment outcome and treatment assignment, simultaneously. In this way, FSRM could mitigate selection bias and minimize the influence of irrelevant variables by simultaneously predicting the treatment assignment and outcomes. FSRM contains deep feature selection, balanced representation learning, and deep prediction network. Deep feature selection uses a sparse one-to-one layer and deep structures to model non-linearity, selecting a subset of features from the input observational data. The deep prediction network helps learn latent representations that are predictive of treatments and observed outcomes. Due to the treatment assignment bias, there is imbalance between the original treatment and control distributions. Matching is not always perfect due to incomplete overlap between treatment and control groups, which will lead to biased estimates of treatment effects. To address this issue, the balanced representation learning component of FSRM attempts to reduce the discrepancy between the two distributions by incorporating a regularizer based on the Wasserstein distance (Sriperumbudur et al., 2012). Finally, FSRM performs matching in the balanced representation space to estimate treatment effects. To the best of our knowledge, FSRM is the first matching method which seamlessly integrates deep feature selection and deep representation learning for causal inference together with joint prediction of treatment assignment and counterfactual outcomes for causal inference. We evaluate the proposed FSRM method on IHDP, modified IHDP and simulated datasets, and demonstrate its superiority over the state-of-the-art methods for treatment effects estimation, especially when the data includes different types of variables.

We organize the rest of our thesis as follows. Technical background including the basic notations, definitions, and assumptions are introduced in Chapter 2. Chapter 3 reviews related work. Our proposed framework is presented in Chapter 4. In Chapter 5, experiments on IHDP, modified IHDP and simulation dataset are provided. We close with a conclusion in Chapter 6.

Chapter 2

Background

Suppose that the observational data contain n units, and that each unit received one of two or more treatments. Let t_i denote the treatment assignment for unit i; $i = 1, ..., n$. For binary treatments, $t_i = 1$ for the treatment group, and $t_i = 0$ for the control group. The outcome for unit i is denoted by Y_t^i when treatment t is applied to unit i; that is, Y_1^i is the potential outcome of unit i in the treatment group and Y_0^i is the potential outcome of unit i in the control group. For observational data, only one of the potential outcomes is observed according as the actual treatment assignment of unit i. The observed outcome is called the factual outcome and remaining unobserved potential outcomes are called counterfactual outcomes. Let $X \in \mathbb{R}^d$ denote all observed variables of a unit.

In this thesis, we follow the potential outcome framework for estimating treatment effects (Rubin, 1974; Splawa-Neyman et al., 1990). The individual treatment effect (ITE) for unit i is the difference between the potential treated and control outcomes, and is defined as:

$$\text{ITE}_i = Y_1^i - Y_0^i, \quad (i = 1, ..., n). \tag{2.0.1}$$

The average treatment effect (ATE) is the difference between the mean potential treated and control outcomes, which is defined as:

$$\text{ATE} = \frac{1}{n}\sum_{i=1}^{n}(Y_1^i - Y_0^i), \quad (i = 1, ..., n). \tag{2.0.2}$$

The success of the potential outcome framework is based on the following assumptions (Imbens & Rubin, 2015), which ensure that the treatment effect can be identified.

Stable Unit Treatment Value Assumption (SUTVA): The potential outcomes for any units do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. **Consistency**: The potential outcome of treatment T is equal to the observed outcome if the actual treatment received is T. **Positivity**: For any value of X, treatment assignment is not deterministic, i.e.,$P(T = t|X = x) > 0$, for all t and x. **Ignorability**: Given covariates X, treatment assignment T is independent to the potential outcomes, i.e., $(Y_1, Y_0) \perp\!\!\!\perp T|X$.

Chapter 3

Related Work

Embracing the rapid developments in machine learning and deep learning, various causal effect estimation methods for observational data have sprung up. Balancing neural networks (BNNs) (F. Johansson et al., 2016) and counterfactual regression networks (CFRNET) (Shalit et al., 2017) are proposed to balance covariate distributions across treatment and control groups by formulating the problem of counterfactual inference as a domain adaptation problem. This model is extended to any number of treatments even with continuous parameters, as described in the perfect match (PM) approach (Schwab et al., 2018) and DR-Nets (Schwab et al., 2019). Following this idea, a few improved models have been proposed and discussed. For example, the shift-invariant representation learning is combined with re-weighting methods (F. D. Johansson et al., 2018). A local similarity preserved individualized treatment effect (SITE) estimation method (Yao et al., 2018) is proposed focusing on local similarity information that provides meaningful constraints on individual treatment estimation. Generative adversarial networks (Yoon et al., 2018) for individualized treatment effects have also been proposed for individual treatment effect estimation.

Besides the popular representation learning methods, matching methods are also among the mostly widely used approaches to causal inference from observational data. The core

purpose of matching methods is to reduce the estimation bias brought on by confounders, so how to find the most similar neighbors in opposite treatment group is the most important problem. The similarity among neighbours can be measured by different distance metrics, including distance based on original covariates space such as the Euclidean distance (Rubin, 1973) and Mahalanobis distance (Rubin & Thomas, 2000), and distance based on transformed space such as the propensity score (Rosenbaum & Rubin, 1983), prognosis score (Hansen, 2008), random subspaces (S. Li et al., 2016), and balanced and nonlinear representation based nearest neighbor matching (BNR-NNM) (S. Li & Fu, 2017).

We propose a feature selection representation matching method in this thesis, which inherits the advantages of both the matching based methods and the representation learning based methods. Different from existing work on treatment effect estimation, our method learns a selective, nonlinear, and balanced representation through deep neural networks, and performs matching in the representation space. Incorporating feature selection layers into deep representation learning, which simultaneously predicts the treatment assignment and outcomes, makes the representation space best predictive of individual treatment outcome, mitigate treatment selection bias, and minimize the influence of irrelevant variables.

Chapter 4

The Proposed Framework

4.1 Motivation

Estimating treatment effects from observational data is faced with two major challenges, i.e., missing counterfactual outcomes and treatment selection bias. To overcome these challenges, we aim to propose a new matching method with the following characteristics:

Selective. If all observed variables are directly used to estimate treatment effects, more bias will be introduced into the model as discussed in chapter 1. To address this issue, it is critical to select important features from observational data. Such a feature selection mechanism will also make the deep neutral networks and matching estimators more interpretable.

Nonlinear. Observational data always involve a large number of variables with complicated relationships among them. Under deep neural networks, the data inform the relationship between outcomes and predictors with enough flexibility to describe complicated data distributions. Therefore, deep neural networks could be used to learn nonlinear representations for treated and control units, and then benefit the matching procedure.

Adjustable. The dimension of the representation vector is adjustable according to complexity of data. It can avoid the curse of dimensionality caused by matching in the high-

dimensional original covariate space or information loss caused by one-dimensional propensity scores.

Balanced. Like propensity score matching by controlling for the covariates that predict treatment assignments, matching in the representation space that is predictive of treatment assignments can also reduce the selection bias due to confounding variables. In addition, matching will lead to biased estimates of treatment effects due to incomplete overlap between treatment and control groups. We incorporate Wasserstein distance to measure distance between representation distributions of treatment and control groups to ensure that the Positivity assumption holds true.

Individualized. Most matching methods can only perform well for estimating the average treatment effect but have poor performance in individual treatment effect estimation compared with regression-based methods. The main reason is that matching methods mainly focus on reducing the bias due to confounding variables but neglect the prediction of observed outcomes based on pre-treatment variables. Our representation space can, by predicting observed outcomes, best represent pre-treatment variables, which means that our matching method has competitive performance with respect to individual treatment effect estimation.

4.2 Model Architecture

We propose a feature selection representation matching (FSRM) method based on deep representation learning and matching in the representation space. The key idea of FSRM is to map the original covariate space into a selective, nonlinear, and balanced representation space, which can be best predictive of individual treatment outcomes, mitigate selection bias, and minimize the influence of irrelevant variables by simultaneously predicting the treatment assignment and outcomes. The framework of FSRM is illustrated in Fig. 4.1, which contains five major components: feature selection layer, deep representation layers,

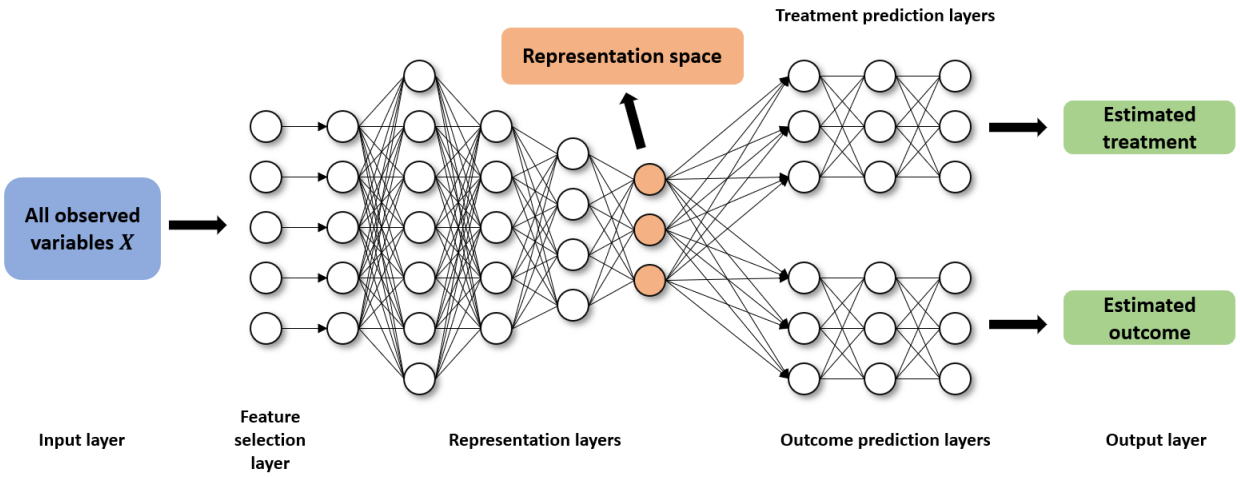

Figure 4.1: The framework of the proposed feature selection representation matching.

outcome prediction layers, treatment prediction layers, and matching in the representation space.

The proposed FSRM method uses the learned representation vectors as the balancing scores to perform matching. The first stage of FSRM adopts a feature selection layer, which is expressed as a nonlinear mapping $\Phi : X \to R$, where X denotes the original covariate space and R denotes the representation space. In the second stage, FSRM not only predicts the treatment assignment through the function h_1, which maps R onto the treatment space $T = \{0, 1\}$, but also predicts the observed outcome through the function h_2, which maps $R \times T$ onto the outcome space Y.

Moreover, to satisfy the requirement of matching method that treatment and control groups overlap with respect to covariates, we need to balance the representation distributions between the two groups. The Integral Probability Metric (IPM) (Shalit et al., 2017; Sriperumbudur et al., 2012) is incorporated into FSRM to maximize the overlap of represen-

tation distributions of treatment and control groups. Finally, optimal matching is performed on the selective, nonlinear, and balanced representations.

Let $\Phi(x)$, $h_1(\Phi(x))$ and $h_2(\Phi(x), t)$ be parameterized by deep neural networks trained jointly in an end-to-end fashion. Deep neural networks are models structured by multiple hidden layers with non-linear activation functions. They can often dramatically increase prediction accuracy, describe complex relationships, and generate structured high-level representation of features which can assist interpretation of data. In the following, we introduce each component of FSRM in detail.

4.2.1 Deep Feature Selection

For $\Phi : X \to R$, we adopt a deep feature selection model (Y. Li et al., 2016) that enables variable selection in deep neural networks. This model takes advantage of deep structures to capture data non-linearity and conveniently selects a subset of features of the data at the input level and following representation layers. In this model, the first feature selection layer is a sparse one-to-one layer between the input and the first hidden layer. Feature selection at the input level can help select which variables are input into the neural network and used for representing pre-treatment variables, which makes the deep neutral network more interpretable.

In the first feature selection layer, every input variable only connects to its corresponding node where the input variable is weighted. This is a 1-1 layer instead of fully connected layer. To select input features, weights w in the feature selection layer and the following representation layers have to be sparse and only the features with nonzero weights are selected to enter the following layers.

LASSO (Tibshirani, 1996) was considered first for this purpose. It is a penalized least squares method imposing the L_1-penalty on the regression coefficients by $\Re(w) = \|w\|_1$.

However, for observational data with high dimensional variables, LASSO cannot remove enough variables before it saturates. To overcome this limitation, the elastic net (Zou & Hastie, 2005) is adopted in our model, which adds a quadratic term $\|w\|_2^2$ to the penalty, that is $\Re(w) = \lambda\|w\|_2^2 + \alpha\|w\|_1$, where λ and α are trade-off parameters. Therefore, for feature selection in the deep neural networks of FSRM, we minimize the objective function:

$$\Re = \lambda \sum_{s=1}^{S} \|w^{(s)}\|_2^2 + \alpha \sum_{s=1}^{S} \|w^{(s)}\|_1, \tag{4.2.1}$$

where S is the number of hidden layers including the feature selection layer and the representation layers in FSRM. The terms $\lambda \geq 0$ and $\alpha \geq 0$ are hyper-parameters that not only control the trade-off between regularization term and the following objective terms, but also control the trade-off between smoothness and sparsity of the weights in the feature selection layer (Y. Li et al., 2016).

As discussed in this chapter, we combine two ideas: a sparse one-to-one feature selection layer between the input and the first hidden layer selects which variables are input into the neural network and elastic net throughout the fully-connected representation layers assigns larger weights to important features. This strategy can effectively filter out the irrelevant variables and highlight the important variables.

4.2.2 Deep Prediction Network

For $h_1(\Phi(x))$ and $h_2(\Phi(x), t)$, we adopt two branches of deep neural networks to predict the outcomes Y_i and treatment assignments T_i based on the representations $\Phi(x)$, as illustrated in Fig. 4.1. Each branch is implemented by fully connected layers and one output regression layer.

The function $h_1(\Phi(x))$ maps the representation vector to the corresponding observed treatment assignment T_i. We use the cross entropy loss $\mathcal{L}_T$ to quantify the factual treatment prediction error:

$$\mathcal{L}_T = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}(t_{ij}log(\hat{t}_{ij})), \tag{4.2.2}$$

where i indexes units and j indexes the treatment assignment classes. The terms t_{ij} are the factual probability distributions over K classes for unit i for each treatment assignment j, and $\hat{t}_{ij}$ are the predicted probability distributions over K classes for unit i. A squared L_2 norm regularization term, $\beta\sum_{p=1}^{P}\|w^{(p)}\|_2^2$, is placed on the model parameters $w^{(p)}$ to mitigate over-fitting, where $\beta \geq 0$ denotes a tuning constant controlling the trade-off between the L_2 regularization term of prediction layers and other terms in final objective function.

The function $h_2(\Phi(x), t)$ maps the representation vector and treatment assignment to the corresponding observed outcome. However, when the dimensionality of representation space is high, there is a risk of losing the influence of t on $h_2(\Phi(x), t)$ if the concatenation of $\Phi(x)$ and t is treated as input (Shalit et al., 2017). To solve this issue, $h_2(\Phi(x), t)$ is partitioned into two head layers $h_2^0(\Phi)$ and $h_2^1(\Phi)$:

$$h_2(\Phi(x), t) = \begin{cases} h_2^0(\Phi) & \text{if } t = 0 \\ h_2^1(\Phi) & \text{if } t = 1. \end{cases} \tag{4.2.3}$$

Here, the first layer $h_2^1(\Phi)$ is used to estimate the outcome under treatment and the second layer $h_2^0(\Phi)$ is used to estimate the outcome for the control group. Each sample is only updated in the head layer corresponding to the observed treatment. Obviously, this model can also be extended to any number of treatments. Let $\hat{y}_i = h_2(\Phi(x), t)$ denote the inferred observed outcome of unit i corresponding to factual treatment t_i. We aim to minimize the

mean squared error in predicting factual outcomes

$$\mathcal{L}_Y = \delta \frac{1}{n} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2, \tag{4.2.4}$$

where the hyper-parameter δ controls the trade-off between the outcome prediction and treatment prediction loss functions. A squared L_2 norm regularization term on the model parameters, $\beta \sum_{p=1}^{P} \|w^{(p)}\|_2^2$, is added to mitigate the overfitting problem, where $\beta \geq 0$ denotes the hyper-parameter controlling the trade-off between the L_2 regularization term of the prediction layers and other terms in the final objective function.

4.2.3 Learning Balanced Representations

The deep feature selection and deep prediction networks learn compact and nonlinear representations for control and treated units. However, the distributions of the treatment group and control group might be imbalanced in the representation space. For the matching procedure, the representation distributions of treatment and control groups should have overlap. To this end, we adopt integral probability metrics (IPM) when learning the representation space to make sure that the nonlinear representation distributions are balanced for the two groups. The integral probability metrics measure the divergence between the representation distributions of treatment and control groups, so we want to minimize the IPM to make two distributions more similar. Let $P(\Phi(x)|t = 1)$ and $Q(\Phi(x)|t = 0)$ denote the empirical distributions of the representation vectors for the treatment and control groups, respectively. In FSRM, we adopt the IPM defined in the family of 1-Lipschitz functions, which leads to IPM being the Wasserstein distance (Guo et al., 2019; Shalit et al., 2017; Sriperumbudur et al., 2012). In particular, the IPM term with Wasserstein distance is defined as

$$\text{Wass}(P, Q) = \gamma \inf_{k \in \mathcal{K}} \int_{\Phi(x)} \|k(\Phi(x)) - \Phi(x)\| P(\Phi(x)) d(\Phi(x)), \tag{4.2.5}$$

where γ denotes the hyper-parameter controlling the trade-off between Wass(P, Q) and other terms in the final objective function. $\mathcal{K} = \{k|Q(k(\Phi(x))) = P(\Phi(x))\}$ defines the set of push-forward functions that transform the representation distribution of the treatment distribution P to that of the control Q and $\Phi(x) \in \{\Phi(x)_i\}_{i:t_i=1}$.

4.2.4 Objective Function

Putting all the above together, the objective function of our feature selection representation matching (FSRM) model is:

$$\begin{aligned}\mathcal{L} = &-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}(t_{ij}log(\hat{t}_{ij})) \\ &+ \delta\frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i^{t_i} - y_i)^2 \\ &+ Wass(P, Q) \\ &+ \lambda\sum_{s=1}^{S}\|w^{(s)}\|_2^2 + \alpha\sum_{s=1}^{S}\|w^{(s)}\|_1 \\ &+ \beta\sum_{p=1}^{P}\|w^{(p)}\|_2^2.\end{aligned} \tag{4.2.6}$$

The first term in the expression above is the loss function for factual treatment assignment prediction. The second is the loss function for prediction of observed outcomes. The third term is to balance representation distributions of treatment and control group. Fourth term is elastic net term, used for deep feature selection and regularization. The last term regularizes the deep prediction network. By minimizing this objective function, FSRM obtains a nonlinear and balanced representation space, which can best predict individual treatment

outcomes, mitigate selection bias, and ignore as much as possible the irrelevant variables. FSRM is implemented using standard feed-forward neural networks with Dropout (Srivastava et al., 2014) and the ReLU activation function. Adam (Kingma & Ba, 2014) is adopted to optimize the objective function.

4.2.5 Matching in Representation Space

Leveraging the selective, balanced, and nonlinear representations extracted from observational data, we perform optimal matching (Rosenbaum, 1989) to find the matched samples with the smallest average absolute distance across all the matched pairs. In our work, the distance between treatment and control units is calculated based on Euclidean distance, Mahalanobis distance, and propensity score, respectively. The outcome of the selected control (treatment) unit within matched pair serves as the estimated counterfactual of the corresponding treatment (control) unit within each matched pair.

Chapter 5

Experiments

In this Chapter, we conduct experiments on three datasets, including the IHDP, modified IHDP, and a synthetic dataset, to evaluate the following aspects: (1) Our proposed method can improve treatment effect estimation with respect to average treatment effect and individual treatment effect. (2) The deep feature selection layers can help improve the performance of treatment effect estimation from observational data with high-dimensional variables or in the presence of different types of variables. (3) The proposed model is robust to different levels of treatment selection bias.

5.1 Datasets

IHDP. The IHDP dataset is a commonly adopted benchmark collected by the Infant Health and Development Program (Brooks-Gunn et al., 1992). These data are generated based on a randomized controlled trial where intensive high-quality care and specialist home visits were provided to low-birthweight and premature infants. There are a total of 25 pre-treatment covariates and 747 units, including 608 control units and 139 treatment units. The outcome is the infants' cognitive test score which can be simulated using the pre-treatment covariates

and the treatment assignment information through the NPCI package [1]. In the IHDP, a biased subset of the treatment group is removed to simulate the selection bias (Shalit et al., 2017). We repeat these procedures 1000 times to conduct evaluations of uncertainty of estimates.

Modified IHDP. In practice, we cannot be sure that all collected variables are always relevant to the study. These variables can bring about extra uncertainty and bias. For IHDP, the outcome is simulated based on all the pre-treatment covariates in the IHDP dataset, so there are no irrelevant variables. Many observational studies include irrelevant variables that are related to neither the treatment nor the outcome of interest. To mimic this situation, we added 35 irrelevant variables as noise to increase data complexity. These additional variables, which are not associated with either the outcomes or the treatment assignments, are sampled from a multivariate normal distribution with mean 0 and random positive definite covariance matrix based on a uniform distribution over the space 35×35 of the correlation matrix (Jacob et al., 2019). Compared to the original IHDP dataset, treatment effect estimation for the modified IHDP dataset is more challenging due to the irrelevant variables.

Synthetic Dataset. To mimic situations where large numbers of variables and information on instrumental, adjustment, confounding, and irrelevant variables are available, we generate a synthetic dataset which reflects the complexity of observational medical records data. Our synthetic data includes confounders, instrumental, adjustment and irrelevant variables. The interrelations among these variables, treatments, and outcomes are illustrated in Fig. 5.1. The number of observed variables in the vector $X = (C^\intercal, Z^\intercal, I^\intercal, A^\intercal)^\intercal$ is set to 60, including 15 confounders in C, 15 adjustment variables in A, 10 instrumental variables in Z, and 20 irrelevant variables in I. The model used to generate the continuous outcome variable Y

[1] https://github.com/vdorie/npci

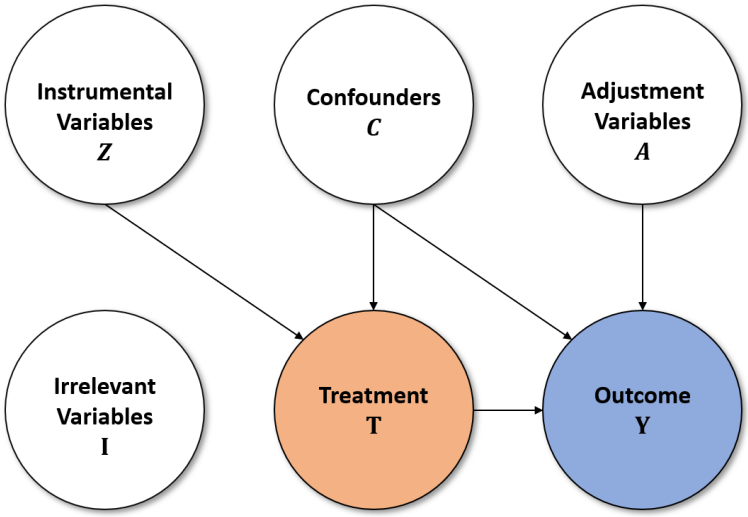

Figure 5.1: The types of observed variables.

Table 5.1: Performance comparison on IHDP, modified IHDP and synthetic data.

	IHDP		modified IHDP		synthetic data	
Method	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
kNN	4.10 ± 0.20	0.79 ± 0.05	4.84 ± 0.32	0.84 ± 0.07	0.40 ± 0.04	0.06 ± 0.05
CF	3.80 ± 0.20	0.40 ± 0.03	4.02 ± 0.36	0.47 ± 0.06	0.38 ± 0.02	0.05 ± 0.03
RF	6.60 ± 0.30	0.96 ± 0.06	7.23 ± 0.45	1.02 ± 0.07	0.37 ± 0.02	0.04 ± 0.03
BART	2.30 ± 0.10	0.34 ± 0.02	3.65 ± 0.21	0.48 ± 0.03	0.38 ± 0.02	0.07 ± 0.04
GANITE	2.40 ± 0.40	0.49 ± 0.05	3.71 ± 0.43	0.54 ± 0.05	0.48 ± 0.05	0.07 ± 0.05
PSM	2.70 ± 3.85	0.49 ± 0.81	4.15 ± 3.21	1.35 ± 0.75	0.48 ± 0.06	0.20 ± 0.13
TARNET	0.95 ± 0.02	0.28 ± 0.01	2.23 ± 0.10	0.48 ± 0.03	0.41 ± 0.02	0.06 ± 0.04
$\text{CFRNET}_{\text{wass}}$	0.76 ± 0.02	0.27 ± 0.01	1.95 ± 0.08	0.38 ± 0.02	0.39 ± 0.03	0.06 ± 0.04
SITE	$\mathbf{0.66 \pm 0.11}$	0.20 ± 0.01	1.88 ± 0.14	0.36 ± 0.02	0.37 ± 0.03	0.05 ± 0.04
PM	0.84 ± 0.61	0.24 ± 0.01	2.12 ± 0.53	0.45 ± 0.03	0.37 ± 0.04	0.06 ± 0.05
$\text{FSRM}_{\text{Mahal}}$	1.32 ± 0.05	0.07 ± 0.01	1.39 ± 0.08	0.07 ± 0.02	0.15 ± 0.01	$\mathbf{0.01 \pm 0.01}$
$\text{FSRM}_{\text{Euclid}}$	1.24 ± 0.04	$\mathbf{0.05 \pm 0.01}$	$\mathbf{1.31 \pm 0.10}$	$\mathbf{0.06 \pm 0.01}$	$\mathbf{0.13 \pm 0.01}$	$\mathbf{0.01 \pm 0.01}$
$\text{FSRM}_{\text{Propensity}}$	3.05 ± 0.15	0.09 ± 0.01	3.09 ± 0.42	0.09 ± 0.02	0.20 ± 0.01	$\mathbf{0.01 \pm 0.01}$

in this simulation is the partially linear regression model (Eq. (5.1.1)) extending the ideas described in (Jacob et al., 2019; Robinson, 1988):

$$Y = \tau((C^\intercal, A^\intercal)^\intercal)T + g((C^\intercal, A^\intercal)^\intercal) + \epsilon, \tag{5.1.1}$$

where $T \overset{ind.}{\sim} \text{Bernoulli}(e_0((C^\intercal, Z^\intercal)^\intercal))$. ϵ are unobserved covariates, which follow a random standard normal distribution $N(0,1)$ and $E[\epsilon|C, A, T] = 0$.

We generate the confounders $C \in \mathbb{R}^{15}$, adjustment variables $A \in \mathbb{R}^{15}$, instrumental variables $Z \in \mathbb{R}^{10}$, and irrelevant variables $I \in \mathbb{R}^{20}$ in a way the variables in each type of variables are partially correlated among each other. They are generated by multivariate normal distribution with mean 0 and random positive definite covariance matrix based on a uniform distribution over the space 15×15, 15×15, 10×10, 20×20 of the correlation matrix, respectively (Jacob et al., 2019). The function $\tau((C^\intercal, A^\intercal)^\intercal)$ describes the true treatment effect as a function of the values of adjustment variables A and confounders C; namely $\tau((C^\intercal, A^\intercal)^\intercal) = (\sin((C^\intercal, A^\intercal)^\intercal \times b_\tau))^2$ where b_τ represents weights for every covariate in the function, which is generated by uniform$(0,1)$. The variable treatment effect implies that its strength differs among the units and is therefore conditioned on C and A. The function $g((C^\intercal, A^\intercal)^\intercal)$ can have an influence on outcome regardless of treatment assignment. It is calculated via a trigonometric function to make the covariates non-linear, which is defined as $g((C^\intercal, A^\intercal)^\intercal) = (\cos((C^\intercal, A^\intercal)^\intercal \times b_g))^2$. Here, b_g represents a weight for each covariate in this function, which is generated by uniform$(0,1)$. The bias is attributed to unobserved covariates which follow a random normal distribution $N(0,1)$. The treatment assignment T follows the Bernoulli distribution, i.e., $T \overset{ind.}{\sim} \text{Bernoulli}(e_0((C^\intercal, Z^\intercal)^\intercal))$ with probability $e_0((C^\intercal, Z^\intercal)^\intercal) = \Phi(\frac{a-\mu(a)}{\sigma(a)})$, where $e_0((C^\intercal, Z^\intercal)^\intercal)$ represents the propensity score, which is the cumulative distribution function for a standard normal random variable based

on confounders C and instrumental variables Z, i.e., $a = \sin((C^\intercal, Z^\intercal)^\intercal \times b_a)$, where b_a is generated by uniform$(0, 1)$.

The total sample size in our synthetic data is 2000, including 1000 units in the treatment group and 1000 units in the control group. During our simulation procedure, $e_0((C^\intercal, Z^\intercal)^\intercal)$ is the propensity score, which represents the treatment selection bias based on their own confounders C and instrumental variables Z. We randomly draw 750 units in the control group and 250 in the treatment group to compose a synthetic dataset with 1000 units. To ensure a robust estimation of model performance, we repeat the random sampling procedure 1000 times and obtain 1000 synthetic datasets.

5.2 Baseline Methods

We compare the proposed feature selection representation matching (FSRM) method with the following baseline methods.

k-nearest neighbor (kNN) method performs matching in the covariate space and use nonparametric preprocessing matching to reduce model dependence in parametric causal inference (Ho et al., 2007). **Causal forests (CF)** is a nonparametric forest-based method for estimating heterogeneous treatment effects by extending Breiman's random forest algorithm (Wager & Athey, 2018). **Random forest (RF)** is a classifier consisting of a combination of tree predictors, in which each tree depends on a random vector that is independently sampled and has the identical distribution for all trees (Breiman, 2001). **Bayesian additive regression trees (BART)** is a nonparametric Bayesian regression model, which uses dimensionally adaptive random basis elements. Every tree in BART model is a weak learner, and it is constrained by a regularization prior. Information can be extracted from the posterior by a Bayesian backfitting MCMC algorithm (Chipman et al., 2010). **Generative adversarial nets for inference of ITE (GANITE)** is based on the Generative

Adversarial Nets framework. It generates proxies of the counterfactual outcomes using a counterfactual generator, and then pass these proxies to an ITE generator (Yoon et al., 2018). **Propensity score matching (PSM)** conducts a matching based on a predicted probability of group membership, which is obtained from logistic regression based on covariates (Ho et al., 2011). **Treatment-agnostic representation network (TARNET)** is a variant of counterfactual regression without the balance regularization (Shalit et al., 2017). **Counterfactual regression ($\mathbf{CFRNET_{wass}}$)** maps the original features into a latent representation space by minimizing the error in predicting factual outcomes and imbalance measured by Wasserstein distance between the treatment representations and the control representations (Shalit et al., 2017). **Local similarity preserved individual treatment effect estimation method (SITE)** is a deep representation learning, which preserves local similarity and balances data distributions simultaneously, by focusing on several hard samples in each mini-batch (Yao et al., 2018). **Perfect match (PM)** augments samples within a minibatch with their propensity-matched nearest neighbours and then implements existing neural network architectures (Schwab et al., 2018).

By using different distance metrics in matching, the proposed FSRM method has three variants denoted as $\text{FSRM}_{\text{Mahal}}$, $\text{FSRM}_{\text{Euclid}}$ and $\text{FSRM}_{\text{Propensity}}$, which adopt the Mahalanobis distance, Euclidean distance, and propensity score, respectively.

5.3 Parameter Settings

The parameters of baseline methods are set the same as suggested in the original papers. To ensure a fair comparison, we follow a standardised approach (Schwab et al., 2018) to hyperparameter optimisation for modified IHDP and the synthetic datasets. The hyperparameters of our method are chosen based on performance on the validation dataset, and the searching range is shown in Table 5.2.

Table 5.2: Hyperparameters and ranges.

Hyperparameter	Range
δ	0, $\{10^k\}_{k=-6}^{2}$, 0.2 , 0.5, 2, 5
γ, λ, α, β	0, $\{10^k\}_{k=-6}^{0}$, 0.2, 0.5
No. and dim. of deep feature selection layers	(dim. of input, 200, 150, 100) (dim. of input, 200, 100, 50) (dim. of input, 100, 100) (dim. of input, 100, 50)
No. of deep prediction layers	1, 2, 3, 4
Dim. of deep prediction layer	50, 100, 150
Batch size	100, 200, 300

5.4 Results and Analysis

For the IHDP, modified IHDP, and our synthetic datasets, we adopt two commonly used evaluation metrics. The first one is the error of ATE estimation, which is defined as:

$$\epsilon_{\text{ATE}} = |\text{ATE} - \widehat{\text{ATE}}|, \tag{5.4.1}$$

where ATE is the true value and $\widehat{\text{ATE}}$ is an estimated ATE.

The second one is the error of expected precision in estimation of heterogeneous effect (PEHE) (Hill, 2011), which is defined as:

$$\epsilon_{\text{PEHE}} = \frac{1}{n}\sum_{i=1}^{n}(\text{ITE}_i - \widehat{\text{ITE}}_i)^2, \tag{5.4.2}$$

where ITE_i is the true ITE for unit i and $\widehat{\text{ITE}}_i$ is an estimated ITE for unit i.

Table 5.1 shows the performance of our method and baseline methods on the IHDP, the modified IHDP, and our synthetic datasets over 1000 realizations. We report the average results and also the standard deviations. FSRM with the Euclidean distance achieves the best performance with respect to ϵ_{ATE} for all three datasets, and the best performance with respect to $\sqrt{\epsilon_{\text{PEHE}}}$ in the modified IHDP and our synthetic dataset. For $\sqrt{\epsilon_{\text{PEHE}}}$ in IHDP, FSRM is better than kNN, CF, RF, BART, GANITE, and PSM, but is outperformed by TARNET, CFRNET, SITE, and PM. Compared with regression-based methods, FSRM is not the top performer, but this result is already remarkable in matching-based methods with respect to the individual treatment effect estimation. In addition, because in the original IHDP data, all of variables are treated as pre-treatment and there are no irrelevant variables, it cannot fully demonstrate the advantages of our method. The inclusion of the feature selection layer in FSRM is not expected to come without cost when all variables are relevant as in the case of IHDP which has no irrelevant variables. This is backed up by the significant gains of FSRM in the modified IHDP dataset in which irrelevant variables are added to IHDP lending extra complexity to the data. Our method can effectively filter out these noises and remain fairly steady in estimating treatment effect with respect to both $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} for the IHDP and modified IHDP. However, the other methods suffered a marked decline in performance for the modified IHDP dataset. Moreover, because our synthetic dataset includes several different types of variables such as instrumental variables, adjustment variables, confounders, and irrelevant variables, our model demonstrates clear superiority over the state-of-the-art methods when dealing with a large number of variables with complicated relationships among them. In addition, we find that $\text{FSRM}_{\text{Euclid}}$ performs better than $\text{FSRM}_{\text{Mahal}}$ and $\text{FSRM}_{\text{Propensity}}$ for all three datasets. Euclidean distance better suits optimal matching based on representation vectors.

5.5 Model Evaluation

Experimental results on three datasets show that FSRM provides more accurate estimation of average treatment effect and individual treatment effect, and it is more highly adaptable to complicated observational data than the state-of-the-art matching estimators and representation learning methods. We further evaluate the performance of FSRM from three perspectives including robustness with respect to different levels of treatment selection bias, the effectiveness of each component of the proposed FSRM, and feature selection interpretability.

Firstly, we evaluate the robustness of our proposed method with respect to different levels of treatment selection bias. Although in our simulation procedure, the treatment selection bias has been taken in account based on their own propensity score $e_0((C^\intercal, Z^\intercal)^\intercal)$, we use conditional sampling from treatment and control groups to increase the treatment selection bias. If the propensity score e_0 is equal to constant 0.5, it means no matter what the confounders and instrumental variables are, the unit is randomly assigned to either the treatment or the control group with the same probability, so that there is no treatment selection bias. The greater $|e_0((C^\intercal, Z^\intercal)^\intercal) - 0.5|$ is, the larger selection bias will end up getting. Following the idea in (Shalit et al., 2017), with probability $1 - q$, we randomly draw the treatment and control units; with probability q, we draw the treatment and control units that have the greatest $|e_0((C^\intercal, Z^\intercal)^\intercal) - 0.5|$. Thus, the higher the q is, the larger the selection bias is. We run $\text{CFRNET}_{\text{WASS}}$, SITE, PM and our method $\text{FSRM}_{\text{Euclid}}$ on the simulation datasets with q from 0 to 1, and show the results in Fig. 5.2 and 5.3. We can observe that our method consistently outperforms the baseline methods under different levels of divergence and is robust to a high level of treatment assignment bias.

In addition, we trained two ablation studies of $\text{FSRM}_{\text{Euclid}}$ on our synthetic dataset. The first one is FSRM (w/o FSL) where the sparse one-to-one feature selection layer between the input and the first hidden layer, and elastic net throughout the fully connected repre-

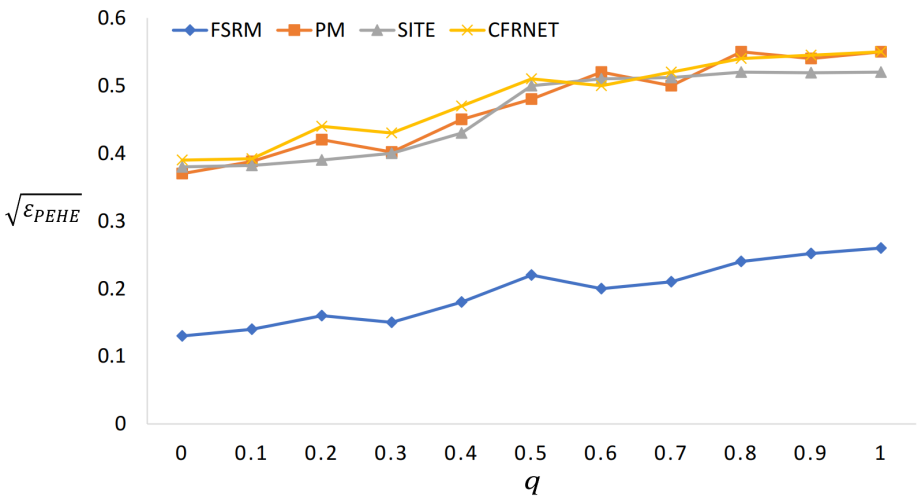

Figure 5.2: $\sqrt{\epsilon_{\text{PEHE}}}$ performance on simulation dataset with q from 0 to 1.

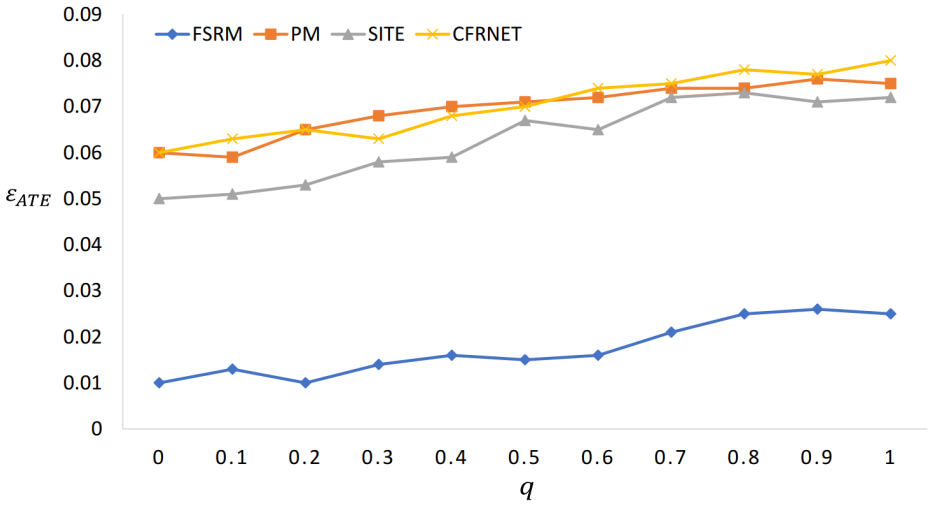

Figure 5.3: ϵ_{ATE} performance on simulation dataset with q from 0 to 1.

Table 5.3: Summary of results in ablation studies.

Synthetic data	Method	$\sqrt{\epsilon_{\text{PEHE}}}$	ϵ_{ATE}
Original bias (q=0)	FSRM	0.13 ± 0.01	0.01 ± 0.01
	FSRM (w/o FSL)	0.41 ± 0.02	0.04 ± 0.01
	FSRM (w/o IPM)	0.15 ± 0.01	0.01 ± 0.01
Extra bias (q=0.5)	FSRM	0.22 ± 0.01	0.02 ± 0.01
	FSRM (w/o IPM)	0.34 ± 0.04	0.03 ± 0.01
Extra bias (q=1)	FSRM	0.26 ± 0.01	0.03 ± 0.01
	FSRM (w/o IPM)	0.58 ± 0.04	0.07 ± 0.03

sentation layers are removed. We only use normal fully connected neural network to learn the representation space. The second ablation study is FSRM (w/o IPM) where the integral probability metric is removed and there is not any restriction on the divergence between the representation distributions of treatment and control groups.

As shown in Table 5.3, the performance becomes poor after removing either the feature selection layers or the IPM module compared to the original FSRM. More specifically, after removing the feature selection layers, $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE} increase dramatically and have similar performance to other baseline methods. Working with the original synthetic data where treatment selection bias is from own propensity score, i.e., $T \overset{ind.}{\sim} \text{Bernoulli}(e_0((C^\intercal, Z^\intercal)^\intercal))$, removing the IPM module from the FSRM (w/o IPM) only has a little impact on $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE}. However if extra bias (q=0.5 and q=1) is added to the synthetic data, the FSRM (w/o IPM) has poor performance compared to original FSRM. In addition, as the extra bias increases, the difference between performance of FSRM (w/o IPM) and performance of original FSRM increases further. Therefore, the feature selection layers and IPM module are essential components of our model.

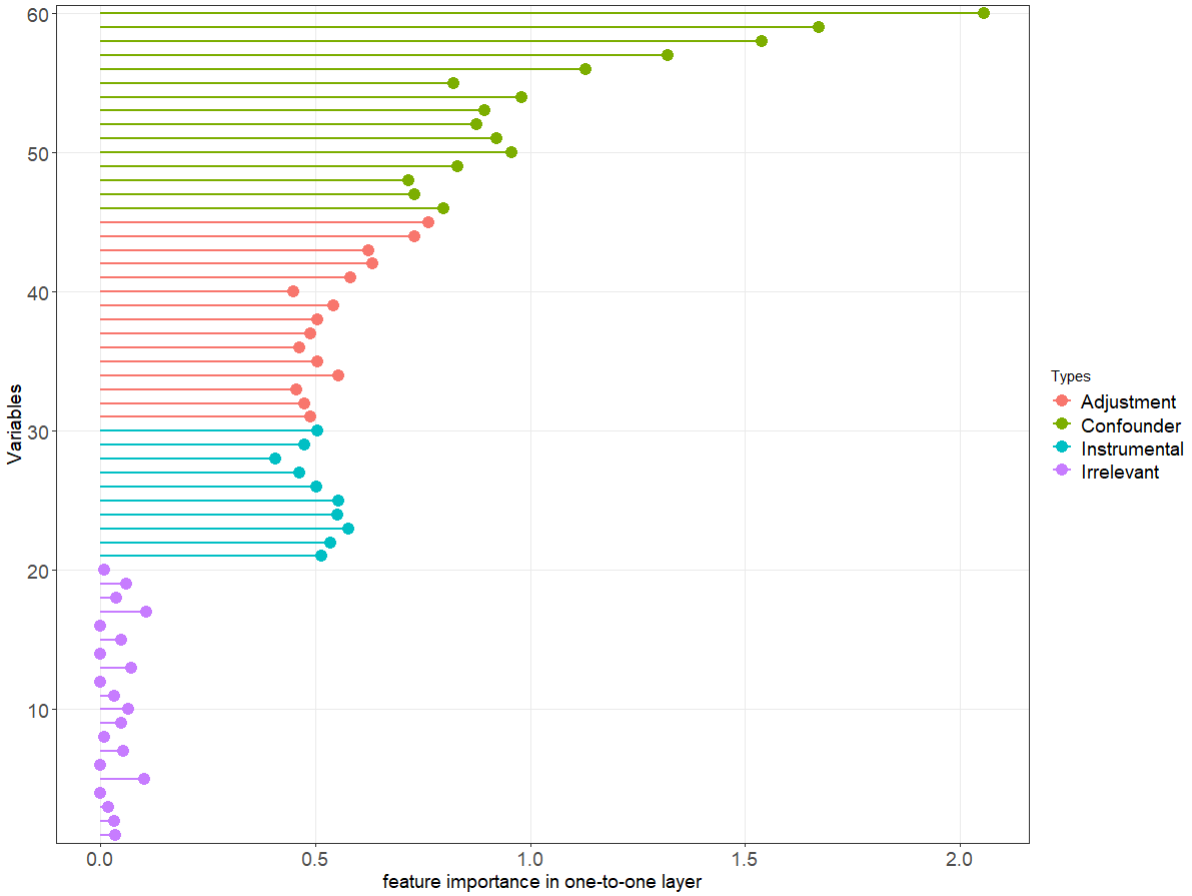

Figure 5.4: Feature importance in one-to-one feature selection layer for 60 input variables.

For feature selection, we have two parts: a sparse one-to-one feature selection layer between the input and the first hidden layer and elastic net throughout the fully connected representation layers. That one-to-one feature selection layer at the input level selects which variables are input into the neural network, which makes the deep neutral network more interpretable. In the one-to-one feature selection layer instead of fully connected layers, every input variable only connects to its corresponding node where the input variable is weighted. This weight can give us an intuitive impression of feature importance for each variable. Fig. 5.4 shows the importance of each feature in the one-to-one layer for the 60 input variables. As we expect, smaller or even zero weights are assigned to irrelevant variables and the largest weights assigned to the confounders. Although the elastic net will continue to conduct feature selection, this figure can give us one clear impression of which variables are forwarded into the following "black box" deeper layers of the deep neural network.

5.6 Sensitivity Analysis

We evaluate FSRM's sensitivity to the three most important parameters δ, λ and α on modified IHDP dataset, which respectively control the weight of observed outcome prediction, smoothness, and sparsity of the weights in the feature selection layer. Based on our analysis presented in Fig.5.5 and 5.6, the performance of our model, in terms of $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE}, is significantly improved compared with the model without L_1 and L_2 penalties. Also, the overall performance on different combinations of hyperparameters of L_1 and L_2 penalties is stable over a large parameter range, which confirms the effectiveness and robustness of deep feature selection in FSRM. This conclusion is consistent with our model evaluation results.

In Fig. 5.7 and 5.8, we find that adding prediction of observed outcome into the model can significantly improve the performance in terms of $\sqrt{\epsilon_{\text{PEHE}}}$ and ϵ_{ATE}, compared with only having the prediction of treatment assignment (i.e., $\delta = 0$). This is the main reason

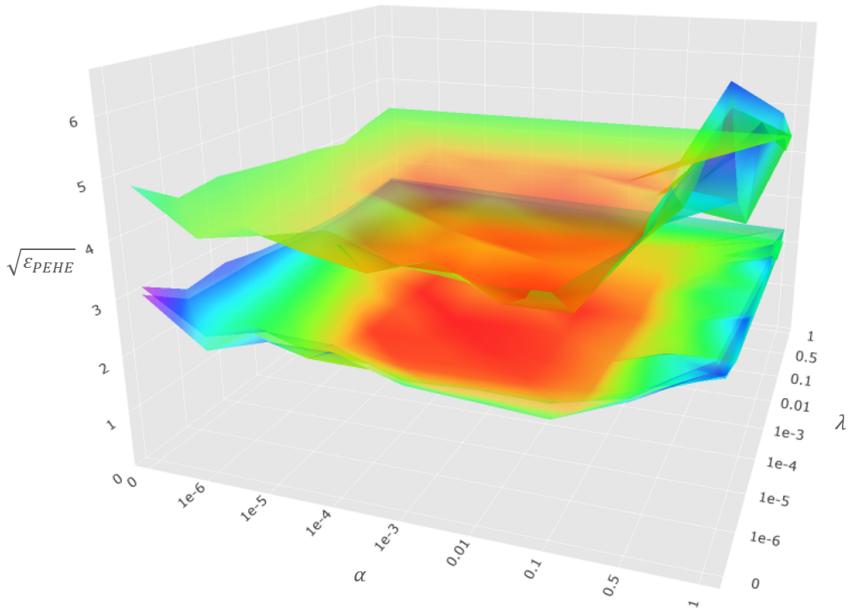

Figure 5.5: $\sqrt{\epsilon_{\text{PEHE}}}$ performance on combinations of hyper-parameters of L_1 and L_2 penalty.

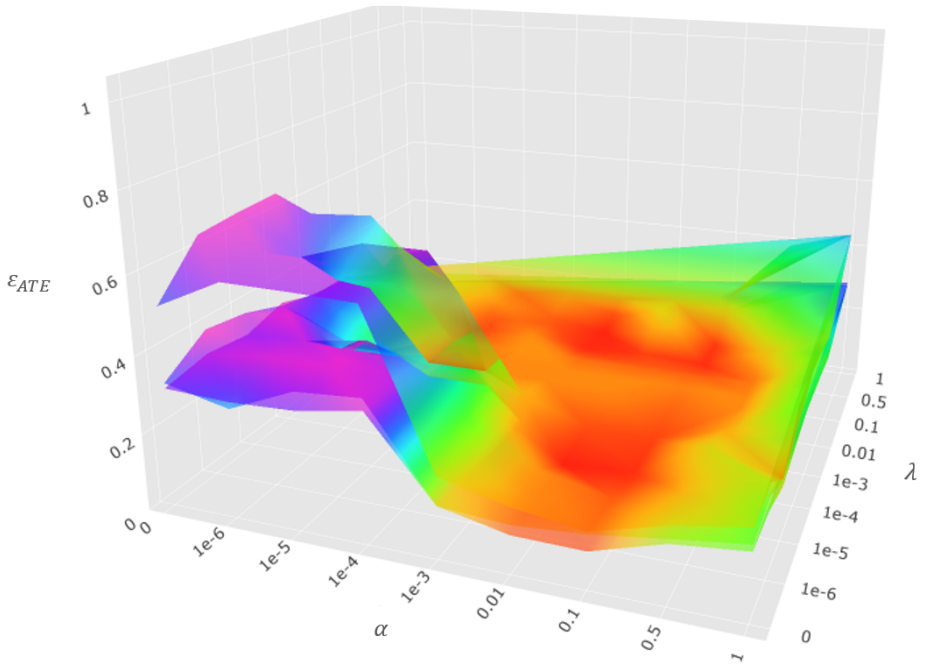

Figure 5.6: ϵ_{ATE} performance on combinations of hyper-parameters of L_1 and L_2 penalty.

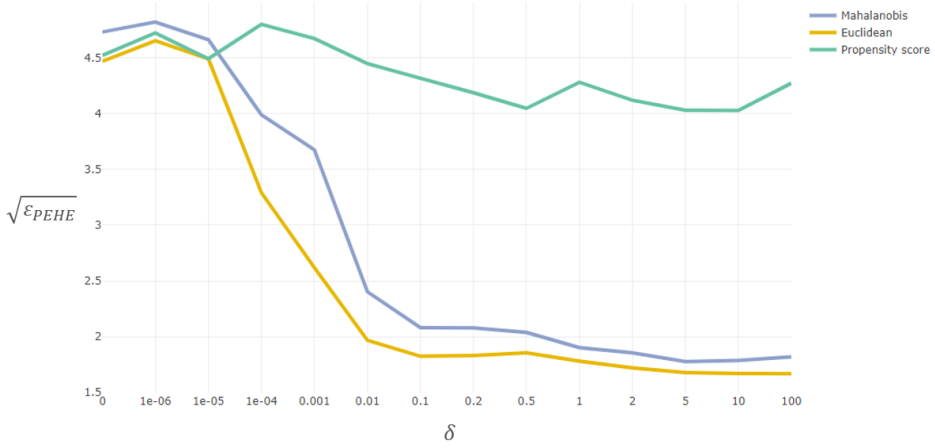

Figure 5.7: $\sqrt{\epsilon_{\text{PEHE}}}$ performance on hyper-parameters δ.

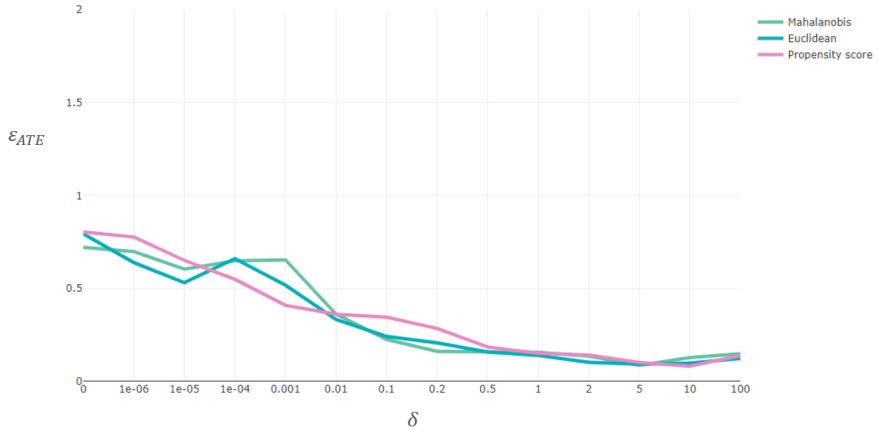

Figure 5.8: ϵ_{ATE} performance on hyper-parameters δ.

why our method performs well when estimating individual treatment effect and average treatment effect, but traditional propensity score matching with logistic regression predicting the treatment assignment cannot accurately estimate individual treatment effects.

Chapter 6

Conclusion

In this thesis, we present a novel feature selection representation matching (FSRM) method for estimating individual treatment effect and average treatment effect, which combines the predictive power of deep learning and interpretability of matching methods. It is applicable to and has good performance for observational data especially with high-dimensional variables or in the presence of different types of variables. Experimental results on three datasets show that FSRM provides more accurate estimation of average treatment effect and individual treatment effect and is more highly adaptable to complicated observational data than the state-of-the-art matching estimators and representation learning methods.

Bibliography

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

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brooks-Gunn, J., Liaw, F.-r., & Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, *120*(3), 350–359.

Chang, Y., & Dy, J. G. (2017). Informative subspace learning for counterfactual inference. *Thirty-First AAAI Conference on Artificial Intelligence*.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Guo, R., Li, J., & Liu, H. (2019). Learning individual treatment effects from networked observational data. *arXiv preprint arXiv:1906.03485*.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, *95*(2), 481–488.

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

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199–236.

Ho, D. E., Imai, K., King, G., Stuart, E. A., et al. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, http://gking. harvard. edu/matchit*.

Iacus, S. M., King, G., & Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Political analysis*, *20*(1), 1–24.

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

Jacob, D., Härdle, W. K., & Lessmann, S. (2019). Group average treatment effects for observational studies. *arXiv preprint arXiv:1911.02688*.

Johansson, F., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. *International conference on machine learning*, 3020–3029.

Johansson, F. D., Kallus, N., Shalit, U., & Sontag, D. (2018). Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, S., & Fu, Y. (2017). Matching on balanced nonlinear representations for treatment effects estimation. *Advances in Neural Information Processing Systems*, 929–939.

Li, S., Vlassis, N., Kawale, J., & Fu, Y. (2016). Matching via dimensionality reduction for estimation of treatment effects in digital marketing campaigns. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3768–3774.

Li, Y., Chen, C.-Y., & Wasserman, W. W. (2016). Deep feature selection: Theory and application to identify enhancers and promoters. *Journal of Computational Biology*, *23*(5), 322–336.

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., & Glynn, R. J. (2011). Effects of adjusting for instrumental variables

on bias and precision of effect estimates. *American journal of epidemiology*, *174*(11), 1213–1222.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*(408), 1024–1032.

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

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, 159–183.

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

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, *95*(450), 573–585.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., & Karlen, W. (2019). Learning counterfactual representations for estimating individual dose-response curves. *arXiv preprint arXiv:1902.00981*.

Schwab, P., Linhardt, L., & Karlen, W. (2018). Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*.

Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3076–3085.

Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 465–472.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G. R., et al. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, *6*, 1550–1599.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2020). A survey on causal inference. *arXiv preprint arXiv:2002.02770*.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., & Zhang, A. (2018). Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 2633–2643.

Yoon, J., Jordon, J., & van der Schaar, M. (2018). GANITE: estimation of individualized treatment effects using generative adversarial nets. *6th International Conference on Learning Representations*.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.