Stable Estimation of Heterogeneous Treatment Effects

Anpeng Wu [1] **Kun Kuang** [1 2] **Ruoxuan Xiong** [3] **Bo Li** [4] **Fei Wu** [1 5 6]

Abstract

Estimating heterogeneous treatment effects (HTE) is crucial for identifying the variation of treatment effects across individuals or subgroups. Most existing methods estimate HTE by removing the confounding bias from imbalanced treatment assignments. However, these methods may produce unreliable estimates of treatment effects and potentially allocate suboptimal treatment arms for underrepresented populations. To improve the estimation accuracy of HTE for underrepresented populations, we propose a novel **Stable** **C**ounter**F**actual **R**egression (**StableCFR**) to smooth the population distribution and upsample the underrepresented subpopulations, while balancing confounders between treatment and control groups. Specifically, StableCFR upsamples the underrepresented data using uniform sampling, where each disjoint subpopulation is weighted proportional to the Lebesgue measure of its support. Moreover, StableCFR balances covariates by using an epsilon-greedy matching approach. Empirical results on both synthetic and real-world datasets demonstrate the superior performance of our StableCFR on estimating HTE for underrepresented populations.

1. Introduction

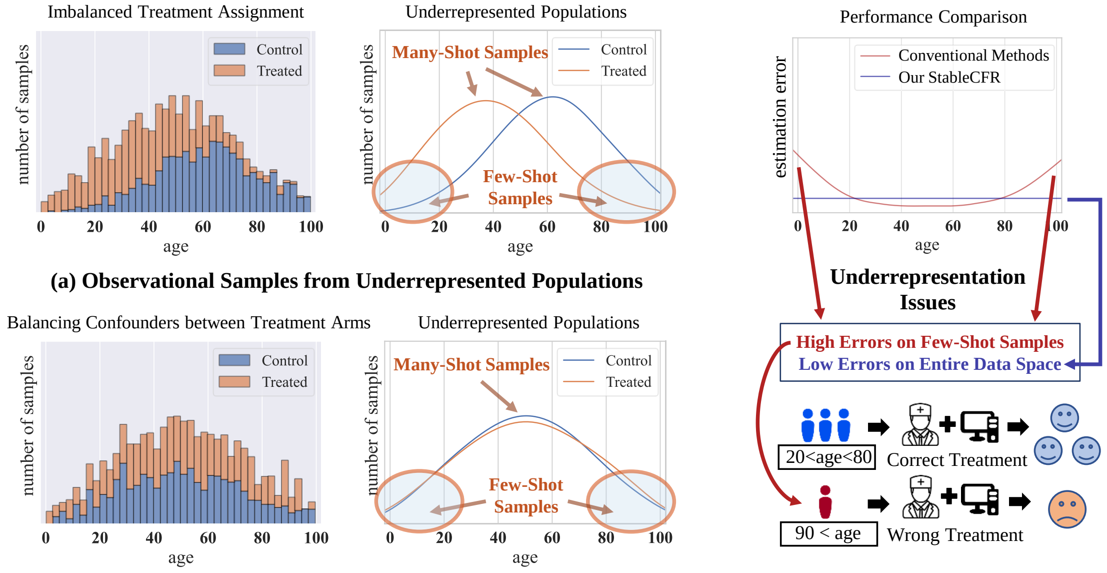
The estimation of heterogeneous treatment effects (HTE) is a crucial problem in causal inference that has been gaining increasing attention across various fields (Imbens et al., 2015; LaLonde, 1986; Pearl, 2009b; Mooij et al., 2016; Shalit et al., 2017), including economics, marketing, biology, and medicine. Non-random treatment assignments in heterogeneous individuals or subgroups can lead to imbalanced confounders between treatment arms, which is known as one of the primary sources of non-causal indirect effects (Shalit et al., 2017; Yao et al., 2020). The bias caused by imbalanced confounders is referred to as confounding bias (Pearl, 2009b). Most of the existing causal methods have been designed to reduce the confounding bias from imbalanced treatment assignments. Although many causal methods have been developed to address confounding bias from imbalanced treatment assignments, they may still be prone to estimation errors stemming from underrepresented populations (Erba et al., 2019; Yang et al., 2021), which include subgroups with notably fewer samples, also known as few-shot samples, compared to other subgroups.

For instance, in the study of the effect of a particular treatment on specific influenza across different age groups, physicians will assign different treatment recommendations (e.g., taking the drug or not) according to the patient's individual circumstances (e.g., age). As shown on the left side in Fig. 1(a), doctors will advise younger people to take the medication more often and advise older people not to take it, and the imbalanced treatment assignments would introduce confounding bias. Besides, the non-random health-seeking behavior of individuals and different age distribution in various regions can lead to issues of sample representativeness in certain sample spaces. As shown on the right side in Fig. 1(a), many-shot samples (representing the majority of the population) aged between 20 and 80 make up 90% of the data, while few-shot samples (representing underrepresented subpopulations such as both younger and older patient groups) are relatively rare, comprising only 10% of the data. Due to limited availability of these few-shot samples, they may be underrepresented in the dataset. This can cause traditional causal models to underestimate the importance of these rare subgroups (see Fig. 1(b)), leading to higher estimation errors (see Fig. 1(c)). Consequently, the estimation of HTE may become unreliable, potentially resulting in suboptimal treatments for worst-case samples.

In this paper, we systematically investigate the primary sources of errors in heterogeneous treatment effect estimation for underrepresented populations. As illustrated in Fig. 1(a), we separate the estimation error of HTE into

[1] Department of Computer Science and Technology, Zhejiang University, China [2] Key Laboratory for Corneal Diseases Research of Zhejiang Province, Zhejiang University, China [3] Department of Quantitative Theory and Methods, Emory University, Atlanta, USA [4] School of Economics and Management, Tsinghua University, Beijing, China [5] Shanghai AI Laboratory, Shanghai, China [6] Shanghai Institute for Advanced Study of Zhejiang University, Shanghai, China. Correspondence to: Kun Kuang <kunkuang@zju.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) Observational Samples from Underrepresented Populations

(b) Conventional Methods for Balancing Confounders

(c) Underrepresentation Issues on Few-Shot Samples

Figure 1. (a) In the HTE estimation, the primary sources of estimation error are confounding bias and underrepresentation issues; (b) Conventional methods for balancing confounders still suffer from underrepresentation issues; (c) Underrepresentation issues might result in existing methods giving unreliable treatment effect estimation and allocating suboptimal treatments for underrepresented populations.

two parts: confounding bias from imbalanced confounders between treatment arms and underrepresentation issues on underrepresented few-shot samples. Although many previous causal models have been developed to remove the confounding bias (Li et al., 2016; Yao et al., 2020; Shalit et al., 2017; Yao et al., 2018; Wu et al., 2022), they still suffer from underrepresentation issues. As illustrated in Fig. 1(b), traditional methods adjust the covariates of different treatment groups into the same distribution; however, there still exists underrepresented subpopulations, resulting in a higher error and unreliable estimates of HTE on few-shot samples, compromising the model's generalization. Therefore, there is a high demand to develop an end-to-end learning framework to simultaneously balance the confounding bias and smooth the underrepresentation distribution for the stable estimation of HTE.

Data imbalance is ubiquitous and prevalent in many real-life applications (Yang et al., 2021). Especially, in biomedical applications, e-markets, and social media, observational datasets are typically constructed by pooling from multiple sources or from certain time periods. This raises concerns about the sample representativeness in some sample spaces. The problem is more pronounced when the heterogeneous treatment effects are of primary interest in precision medicine or target marketing. To this end, we propose a novel Stable CounterFactual Regression (StableCFR[1]) architecture, which utilizes uniform sampling to upsample underrepresented data, assigning weights to each subpopulation proportional to the Lebesgue measure of its support, and incorporates an epsilon-greedy matching module for confounder balancing. Specifically, to jointly balance the confounding bias and smooth the underrepresentation distribution, we design a uniformed nearest neighbor batching (UNNB) training that extends mini-batch training to a multivariate uniform distribution, and an epsilon-greedy matching (EPM) algorithm to match samples within control and treatment groups based on their distance to the random points, which preserves an ideal uniform distribution over each feature and avoids frequent sampling in sparse outliers. The main contributions in this paper are as follows:

- We systematically investigate the estimation error in HTE estimation for underrepresented populations, and firstly separate it into two parts: confounding bias and underrepresentation issues. The conventional causal methods effectively address confounder bias arising from imbalanced confounders, but underrepresented populations limit model's generalizability.

- We propose a novel StableCFR architecture with nearest neighbor batch techniques to balance confounders across treatment groups and smooth the underrepresentation distribution. Our work fills the gap in techniques for HTE estimation with underrepresented populations.

- In synthetic and semi-synthetic data, empirical experiments demonstrate the effectiveness of our algorithm. In addition, with the HTE estimation, StableCFR presents some intuitive explanations for the cardiovascular mortality rate (CMR) study.

[1] The code is available at: https://github.com/anpwu/StableCFR

2. Related Work

2.1. Confounder Balancing

In real-world applications, as shown in Fig. 1(a), imbalanced treatment assignment is a common phenomenon because agents may choose treatments for various reasons, which can introduce confounding bias in observational studies. This has motivated the development of many techniques for addressing confounding bias, such as propensity score methods (Rosenbaum & Rubin, 1983; Rosenbaum, 1987; Li et al., 2016), re-weighting methods (Zubizarreta, 2015; Athey et al., 2018), doubly robust methods (Funk et al., 2011), and the backdoor criterion (Pearl, 2009a). Moreover, representation learning has emerged as an advanced approach to address confounding bias in observational studies. The Counterfactual Regression (CFR) method (Johansson et al., 2016; Shalit et al., 2017) is a pioneering approach that proposes a new form of regularization to learn representations with reduced Integral Probability Metric (IPM) distance between treated and control groups. Building upon this thought, SITE (Yao et al., 2018) simultaneously maintains local similarity and balances the distributions of the representation. CFR-ISW (Hassanpour & Greiner, 2019a) utilizes importance sampling weights to improve the representation. Additionally, DR-CFR (Hassanpour & Greiner, 2019b) and DeR-CFR (Wu et al., 2022) present a disentanglement framework to identify confounders from pre-treatment variables and then balances them.

As illustrated in Fig. 1(b), although these methods have addressed confounding bias and adjust the covariates of different treatment groups to the same distribution, underrepresented subpopulations with significantly fewer observations limit model's generalizability, i.e., underrepresentation issues. To this end, we propose a StableCFR architecture to upsample the underrepresented data using uniform sampling to assign weights to each subpopulation proportional to the Lebesgue measure of the support when balancing confounders between treatment groups.

2.2. Imbalanced Regression

Although the class-wise imbalance is well eliminated (He & Garcia, 2009; Shen et al., 2016; Cui et al., 2019; Hong et al., 2021; Liu et al., 2020; He et al., 2021; Yi et al., 2022; Xiang et al., 2020; Tang et al., 2022), the attribute-wise imbalance still persists and hurts the generalization (Tang et al., 2022) and regression on imbalanced data is relatively under-explored. Building on conventional classification method SMOTE (Chawla et al., 2002), SMOTER (Torgo et al., 2013) interpolates both inputs and targets directly to synthesize samples for rare target regions, and SMOGN (Branco et al., 2017) uses Gaussian noise augmentation. Further work (Branco et al., 2018) ensembles and extends these data pre-processing techniques. Recently, DenseWeight (Steininger et al., 2021) weights data points according to the empirical training distribution through kernel density estimation (KDE). DIR (Yang et al., 2021) re-defines Deep Imbalanced Regression problem and proposes label distribution smoothing (LDS) and feature distribution smoothing (FDS) techniques. Ren et al. (2022) revisits MSE and proposes a Balanced MSE to accommodate the imbalanced training label distribution. Some above deep imbalanced regression methods rely on pre-specified data distributions, and heuristic algorithms require learning the density estimation of the data, which involves tuning multiple parameters and may be computationally intensive.

Sensitivity analysis is a valuable method for assessing the model's robustness (Saltelli, 2002; Imai et al., 2010; VanderWeele & Ding, 2017). However, selecting appropriate sensitivity metrics and conducting multiple experiments would be a challenging and time-consuming task. Additionally, the sensitivity strength is hard to define without imposing priors on the model and population distribution. Building on standard causal assumptions, our StableCFR uses a simple but effective uniform re-sampling to smooth the underrepresentation distribution and assign weights to each subpopulation proportional to the Lebesgue measure of the support.

3. Setup and Estimation

3.1. Problem Setup

In this paper, we aim to estimate the heterogeneous treatment effects of underrepresented populations (Fig. 1(a)) under the unconfoundedness assumption. In the observational data $\mathbb{D} = \{\boldsymbol{x_i}, t_i, y_i\}_{i=1}^n$ where n denotes the sample size; for each unit i, we observe its covariates information $\boldsymbol{x_i} \in \mathcal{X}$ (e.g., age), where $\mathcal{X} \subset \mathbb{R}^d$ and d is the dimension of the observed confounders $\boldsymbol{x_i}$. Besides, we also observe the treatment assignment $t_i \in \mathcal{T}$, where $\mathcal{T} = \{0, 1\}$ denotes a set of treatment options (e.g., {0:placebo, 1:drug}), and observe the corresponding outcome $y_i \in \mathcal{Y}$, where $\mathcal{Y} \subset \mathbb{R}$.

Besides, we denote potential control outcome $Y_i(0) \in \mathcal{Y}$ as a result of choosing control arm $t = 0$, and potential treated outcome $Y_i(1) \in \mathcal{Y}$ as a result of choosing treated arm $t = 1$. Then we define the Heterogeneous Treatment Effect for each unit i and the Average Treatment Effect as:

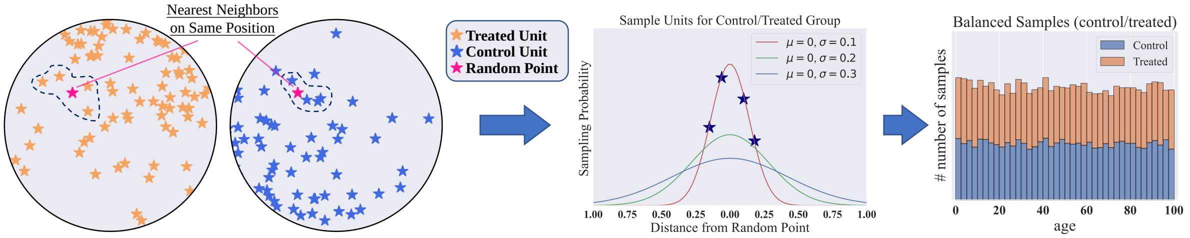
Definition 3.1. Heterogeneous Treatment Effect (HTE):

$$HTE(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}], \tag{1}$$

for simplify, we use HTE_i to denote $HTE(\boldsymbol{x_i})$ of unit i. The HTE is used to estimate the effect of treatment across different subpopulations or groups.

Definition 3.2. Average Treatment Effect (ATE):

$$ATE = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_{\boldsymbol{X}}[HTE(\boldsymbol{X})]. \tag{2}$$



(a) Sample Point from an Ideal Uniform Distribution

(b) Sample Units from Nearest Neighbors

(c) Deep Balance Data

Figure 2. Stable CounterFacutal Regression Architecture.

Our analysis in this paper relies on the following standard assumptions (Imbens et al., 2015) for causal inference.

Assumption 3.3. Stable Unit Treatment Value. The distribution of the potential outcome of one unit is assumed to be independent of the treatment assignment of another unit.

Assumption 3.4. Unconfoundedness. The choice of treatment is independent of the potential outcomes given the covariates. Formally, $T \perp \big(Y(0), Y(1)\big) \mid \boldsymbol{X}$.

Assumption 3.5. Overlap. Each unit has a nonzero probability of being or not being assigned to treatment. Formally, $0 < Pr(T = 1 \mid \boldsymbol{X}) < 1$.

Big datasets in biomedical applications, e-markets, and social media, etc, are typically constructed by pooling from multiple sources or from certain time periods, as shown in Fig. 1. This raises the concern about the sample representativeness in some sample spaces. The underrepresentation issues are more pronounced when the local heterogeneous treatment effects are of primary interest in precision medicine or target marketing. Next, we will introduce the challenges and our solutions to the HTE estimation for underrepresentation populations.

3.2. Stable HTE Estimator

Challenges. As shown in Fig. 1(a), the HTE estimation faces two primary challenges for counterfactual regression. One challenge is imbalanced treatment assignments, which results in different skewed distributions between the control and treated groups, i.e., covariate shift. If we directly perform counterfactual regression on the imbalanced data, it can lead to biased HTE estimation, due to the systemic differences between control and treated groups, i.e., confounding bias. Another challenge is the underrepresentation issues (Fig. 1(c)). Inadequate observations of few-shot samples can cause greater uncertainty and variability in outcome regressions, resulting in higher errors in HTE estimation. Underrepresentation issues might result in existing methods unreliably estimating treatment effects and allocating suboptimal treatments for underrepresented populations.

Although recent conventional causal methods have successfully balanced the distribution of confounders and adjusted them to the same distribution for different treatment groups, as shown in Fig. 1(b), they may suffer from underrepresentation issues due to the limited availability of these underrepresented few-shot samples, leading to unreliable HTE estimation. The underrepresented regression error is still under-explored in causal literature.

Estimator. Regarding the estimation of heterogeneous treatment effects (HTE) using underrepresented data, our concern is that previous regression models prioritize improving the average HTE performance by minimizing mean square error (Eq. (3)). However, this approach may result in unreliable predictions for few-shot samples in underrepresented subpopulations, as noted in Angrist & Pischke (2009). To address this issue, we propose a robust HTE estimator (Eq. (4)) that re-samples the underrepresented data using uniform sampling to assign weights proportional to the Lebesgue measure of the support of each subpopulation.

$$\min \int_{\boldsymbol{x}\in\mathcal{X}} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 dF_X(\boldsymbol{x}), \quad (3)$$

$$\min \int_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{|\mathcal{X}|} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2 d\boldsymbol{x}, \quad (4)$$

where $dF_X(\boldsymbol{x})$ is the density of $\boldsymbol{X}$ in the training data and $|\mathcal{X}|$ denotes Lebesgue measure of the support in $\mathcal{X}$, used to measure the size of support set $\mathcal{X}$ in Euclidean space.

To be more specific, we evenly partition the entire support of X into V disjoint subpopulations, i.e., $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_V$. Then, the objectives corresponding to Eqs. (3) & (4) are as follows:

$$\min \sum_{v=1}^{V} \frac{N_v}{N} \sum_{\boldsymbol{x}\in\mathcal{X}_v} \frac{1}{N_v} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2, \quad (5)$$

$$\min \sum_{v=1}^{V} \frac{|\mathcal{X}_v|}{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}_v} \frac{1}{N_v} \left(\widehat{HTE}(\boldsymbol{x}) - HTE(\boldsymbol{x})\right)^2, \quad (6)$$

where N_v denotes the sample size of $\mathcal{X}_v$, and N denotes the sample size of total training data. Besides, $|\mathcal{X}_v|$ denotes Lebesgue measure of the support in $\mathcal{X}_v$. When there are $V = N$ subpopulations, each subpopulation contains only one sample, i.e., $\mathcal{X}_v = \boldsymbol{x}_v, v = 1, 2, \cdots, N$. In this case, we propose a StableCFR algorithm to achieve Eq. (6) using uniform sampling to assign weights proportional to the Lebesgue measure of the support of each subpopulation.

4. Algorithm

To stable estimate treatment effects for underrepresented populations, as illustrated in Fig. 2, we propose a novel StableCFR architecture to smooth the underrepresentation distribution in few-shot samples, while balancing confounders between treatment and control groups. Specifically, the overall architecture of our StableCFR consists of the following components: (1) **Uniformed Nearest Neighbor Batching (UNNB)** improves mini-batch training by creating batches that follow a multivariate uniform distribution, instead of randomly sampling from the training examples. (2) **Epsilon-greedy Matching (EPM)** can balance the treated and control group and avoid frequent sampling of points in sparse outliers. (3) **One shared representation for two-head prediction networks** $h^0(\phi(\boldsymbol{X}))$ and $h^1(\phi(\boldsymbol{X}))$. The shared representation $\phi(\boldsymbol{X})$ leverages the information from different treatment groups to collaboratively update neural network parameters and learn common patterns among potential outcomes. Next, we will describe each component of our StableCFR algorithm, and go on to further discussion.

4.1. Uniformed Nearest Neighbor Batching

With the advent of deep learning, recent literature (Johansson et al., 2016; Shalit et al., 2017) introduces representation network and uses Integral Probability Metric regularization to learn a balanced representation for counterfactual regression. While these methods have made progress in removing confounding bias and controlling variance, they are prone to focus on many-shot samples and perform poorly on underrepresented few-shot samples, i.e., underrepresentation issues. Furthermore, these methods may learn a overbalanced representation and discard important information related to the predicted treatment variable and confounder information, which may also be predictive of the outcome.

To address the underrepresentation issues and representation overbalancing problem, we upsamples the underrepresented data using uniform sampling to assign weights to each subpopulation proportional to the Lebesgue measure of the support. Thus, under the standard causal assumptions without any additional prior, we propose a nearest neighbor batch to extend mini-batch training to an ideal uniform distribution.

Sample Points from an Ideal Uniform Distribution. Assumption 3.5 implies that the supports of the covariates in the treated and control groups are the same. This motivates us to adjust the imbalanced distributions of the treated and control groups to the same multivariate uniform distribution. We create a multivariate uniform distribution using the supports $\mathcal{X}$ and sample a mini-batch with a batch size of m:

$$\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \cdots, \boldsymbol{q_m}\} \sim \text{Unif}(\boldsymbol{Q}_{min}, \boldsymbol{Q}_{max}), \quad (7)$$

$$\boldsymbol{Q}_{min} = \min(\boldsymbol{X}) - 0.01|\mathcal{X}|, \quad (8)$$

$$\boldsymbol{Q}_{max} = \max(\boldsymbol{X}) + 0.01|\mathcal{X}|, \quad (9)$$

where $\min(\boldsymbol{X})$ and $\max(\boldsymbol{X})$ represent the minimum and maximum values of the covariate $\boldsymbol{X}$ in the supports $\mathcal{X}$, respectively. $|\mathcal{X}|$ represents the support size of the covariates, i.e., $|\mathcal{X}| = \max(\boldsymbol{X}) - \min(\boldsymbol{X})$.

As shown in Fig. 2(a), we draw m random points $\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \cdots, \boldsymbol{q_m}\}$ from multivariate uniform distribution $\text{Unif}(\boldsymbol{Q}_{min}, \boldsymbol{Q}_{max})$, ensuring that any value in the support of covariates has an equal probability of being sampled. This adjusts the original underrepresented population into uniformly distributed data. However, the sampled points $\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \cdots, \boldsymbol{q_m}\}$ are not presented in the training data $\mathbb{D} = \{\boldsymbol{x_i}, t_i, y_i\}_{i=1}^n$, one straightforward strategy is to search their nearest neighbor from training examples.

Match Pairs from Nearest Neighbors. Inspired by covariate matching (Rosenbaum & Rubin, 1983; Abadie & Imbens, 2006; 2011), we search for the two nearest neighbors of the sample point ($\boldsymbol{q_j}, j = 1, 2, \cdots, m$) in the control and treated groups, respectively, and use them as a sample pair $(I_j^{t=0}, I_j^{t=1})$ in the training batch:

$$I_j^{t=0} = \arg\min_{i:t_i=0} \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, \quad (10)$$

$$I_j^{t=1} = \arg\min_{i:t_i=1} \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, \quad (11)$$

where, $\arg\min(\cdot)$ returns the indices of the minimum. $I_j^{t=0}$ denotes the nearest neighbor of point q_j in the control group, and $I_j^{t=1}$ denotes the nearest neighbor in the treated group.

As shown in Fig. 2(a), similar to propensity score matching, we match two nearest neighbors $(I_j^{t=0}, I_j^{t=1})$ of the same point q_j as a sample pair to eliminate the confounding bias. The training batch at each iteration is $\{(I_j^{t=0}, I_j^{t=1})_{j=1,2,\cdots,m}\}$.

4.2. Epsilon-Greedy Matching

Although pure uniform sampling with nearest neighbor samples has balanced the treated and control groups, the greedy strategy of searching for nearest neighbors for matching may lead to frequent sampling of sparse samples in underrepresented populations, hurting the model's performance. Therefore, we select the top-K nearest neighbors and use a hyperparameter ϵ to trade-off exploration (distance-based sampling, with the receptive field controlled by hyperparameter σ) and exploitation (top-K nearest neighbor sampling).

Firstly, we select top-K nearest neighbors of point $\boldsymbol{q_j}$, i.e., $\{I_{j,1}^t, I_{j,2}^t, \cdots, I_{j,K}^t\}$, and perform an epsilon-greedy matching with probabilistic sampling based on the distance, i.e., $D^t_{j,k} = \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, i = I^t_{j,k}$. With probability ϵ, we choose the nearest neighbors of the point $\boldsymbol{q_j}$; with probability $1-\epsilon$, we randomly sample one from the top-K nearest neighbors, and the sampling probability of each neighbor decreases with its distance to q_j. As shown in Fig. 2(b), we use the probability density function $f(\cdot \mid \sigma)$ of the normal distribution $\mathcal{N}(\mu = 0, \sigma)$ as the relative sampling probability in a non-greedy strategy. The closer the sample in training is to the point $\boldsymbol{q_j}$, the higher the probability of the sample being selected for sampling. The Epsilon-Greedy Matching Algorithm, as shown in Alg. 4.2, can be easily extended to multi-value treatment. In this paper, the optimal parameters is $\{K = 10, \epsilon = 0.6, \sigma = 0.25\}$ for both synthetic and semi-synthetic datasets. The discussion about hyper-parameters is deferred to Appendix. B.2.

Algorithm 1 Epsilon-Greedy Matching with Distance-based Sampling for Point q_j.

```
$tmp \sim \text{Unif}(0,1)$
**if** $tmp \leq \epsilon$ **then**
    Return $I^t_{j,1}$.
**else**
    $D^t_{j,k} = \|\boldsymbol{x_i} - \boldsymbol{q_j}\|_2^2, i \in \{I^t_{j,1}, I^t_{j,2}, \cdots, I^t_{j,K}\}$,
    $P^t_{j,k} = f(D^t_{j,k} \mid \sigma), k \in \{1, 2, \cdots, K\}$,
    Sample $I^t_j$ from $\{1, 2, \cdots, K\}$ with relative sampling probability $P^t_{j,k}, k \in \{1, 2, \cdots, K\}$.
    Return $I^t_j$.
**end if**
```

The pure uniform sampling with nearest neighbor sample ($\epsilon = 1$) has already achieved stable counterfactual regression. Furthermore, the uniform sampling with epsilon-greedy matching ($K = 10, \epsilon = 0.6, \sigma = 0.25$) improves HTE estimation and reduces variance by avoiding frequent sampling of sparse samples. With uniform sampling and epsilon-greedy matching, we obtain a balanced matched dataset, as illustrated in Fig. 2(c).

4.3. Neural Network Regression

To stable estimate HTE, our algorithm uses uniform re-sampling and epsilon-greedy matching to ensure that each subpopulation is represented adequately. Additionally, we assign weights proportional to the Lebesgue measure of the support of each subpopulation, and plug the re-sampled balanced data into a crafted two-head neural network. By doing so, we can utilize the benefits of a complex model, while preventing under-fitting in the underrepresented subpopulations. This approach ensures that the performance of the complex model does not decrease significantly, or even only slightly, in the dominant subpopulations.

Shared Backbone. For the shared backbone, we use multi-layer neural network $\phi(\boldsymbol{X})$ with ELU activation function to learn representation and the network has three hidden layers with 128 units, respectively. In the paper, we use the same backbone configuration for all baseline models.

Two-Head Prediction Networks. Then, we adopt two separate neural networks with ELU activation function to predict potential control outcome $h^0(\phi(\boldsymbol{X}))$ and potential treated outcome $h^1(\phi(\boldsymbol{X}))$, and the network has three hidden layers with 128 units, respectively. At each training batch, we use stochastic gradient descent (SGD) to train the network $\{h^0, h^1, \phi\}$ with a loss $\mathcal{L}$ for $10,000$ epochs with a batch size of $m = 100$.

$$\mathcal{L} = \frac{1}{2m}\sum_i \|[t_i \cdot h^1(\phi(\boldsymbol{x_i})) + (1-t_i) \cdot h^0(\phi(\boldsymbol{x_i}))] - y_i\|_2^2,$$

$$i \in \{I_1^{t=0}, I_2^{t=0}, \cdots, I_m^{t=0}, I_1^{t=1}, I_2^{t=1}, \cdots, I_m^{t=1}\}.$$

The same configuration is adopted for all baseline models.

4.4. Discussion

In this paper, the proposed re-sampling strategy guided by multivariate uniform sampling is the core module of the StableCFR algorithm, which is used to address the underrepresentation issue. Instead, the heuristic modification of K-nearest neighbors matching is a sub-module for improving the efficiency of the re-sampling strategy and preventing frequent repetition of sparse samples in underrepresented populations. For heterogeneous treatment effect estimation, we plug the re-sampled balance data into a two-head neural network to automatically learn the counterfactual regression.

However, one limitation of the StableCFR algorithm is the increased computational cost for larger datasets. To mitigate this issue, the solution is to split the large dataset into smaller sub-datasets and randomly select one at a time to create the nearest neighbor batch. Additionally, StableCFR can be used as a pre-processing method, but if the number of batches is too small, the pre-processing may not be able to effectively balance the data because there may not be enough batches to represent all the different subgroups of data. The computational cost is deferred to Appendix. B.3.

5. Experiments

5.1. Datasets

In this paper, we curate five benchmarks with underrepresented populations, including three synthetic datasets, one semi-synthetic and one real-world datasets, to evaluate the effectiveness of our StableCFR in HTE estimation.

Syn-γ ($\gamma = 0.5, 0.8, 1.0$): Under the unconfoundedness assumption, we sample $3,000$ units with an $80/20$ proportion of training/validation splits, and use skewed distributions from underrepresentated populations to generate the covariates $\boldsymbol{X} = \{X_1, X_2, X_3\}$, where X_i denotes i-th variable:

$$\begin{aligned} X_1 &\sim \mathcal{N}(-2.0, 2.0 \mid 0, \gamma), \\ X_2 &\sim \mathcal{N}(-0.1, 2.0 \mid 0, \gamma), \\ X_3 &\sim \mathcal{N}(-2.0, 0.1 \mid 0, \gamma). \end{aligned}$$

We generate three variables $\boldsymbol{X} = \{X_1, X_2, X_3\}$ from the normal distribution, where $\mathcal{N}(a, b \mid 0, \gamma)$ denotes that we sample variable from the range $[a, b]$ of normal distribution $\mathcal{N}(0, \gamma)$ to create different skewed distributions. We expand experiments to various underrepresentation scenarios and adjust the difficulty level of the underrepresentation issues by changing the parameter $\gamma = 0.5, 0.8, 1.0$. The treatment variable T is generated as follows:

$$\begin{aligned} Pr(T \mid \boldsymbol{X}) &= \frac{1}{1 + \exp{(X_1 + X_2 + X_3)/3}}, \\ T &\sim Bernoulli(Pr(T \mid \boldsymbol{X})). \end{aligned}$$

The outcome variable $Y(T)$ is generated as follows:

$$\begin{aligned} &Y(T) = TY(1) + (1-T)Y(0), \\ &Y(0) = |X_2^2 - X_3^2| + 2\cos{(X_1 - X_2 + X_3)} - s(\boldsymbol{X}), \\ &Y(1) = |X_2^2 - X_3^2| + 2\sin{(X_1 - X_2 + X_3)} + s(\boldsymbol{X}), \\ &s(X) = X_1X_2X_3 + X_1^2 + (1 - X_2 + X_3)^2. \end{aligned}$$

Semi-PM-CMR: The PM-CMR (detailed in Appendix A, (Wyatt et al., 2020)) study the impact of $PM_{2.5}$ partical level on the cardiovascular mortality rate (CMR) in $2,132$ counties in the US using the data provided by the National Studies on Air Pollution and Health. As the counterfactual outcomes are rarely available for real-world data, we transfer the $PM_{2.5}$ level in 2010 to generate the treatment variable:

$$\begin{gathered} Pr(T \mid PM_{2.5}) = 0.2 + 0.6 \cdot \frac{PM_{2.5} - \min(PM_{2.5})}{\max(PM_{2.5}) - \min(PM_{2.5})}, \\ T \sim Bernoulli(Pr(T \mid PM_{2.5})), \end{gathered}$$

and we use 7 variables of CMR in 2010 as covariates $\boldsymbol{X} = \{X_1, X_2, \cdots, X_7\}$ to simulate the outcome variable $Y(T)$:

$$\begin{gathered} Y(T) = TY(1) + (1-T)Y(0), \\ Y(0) = \frac{1}{3}\left(\sum_{i=5}^{7} X_i^2 + \sum_{i=3}^{4} X_i X_{i+2}\right) - \sum_{i=1}^{2} X_i + 2\cos{(\sum_{i=1}^{7} X_i)}, \\ Y(1) = \frac{1}{3}\left(\sum_{i=5}^{7} X_i^2 + \sum_{i=3}^{4} X_i X_{i+3}\right) + \sum_{i=1}^{2} X_i + 2\sin{(\sum_{i=1}^{7} X_i)}. \end{gathered}$$

Real-PM-CMR: In this paper, according to the HTE, we also give some intuitive explanations for the relationship between $PM_{2.5}$ partical level and the cardiovascular mortality rate (CMR). In the real application, we use 7 variables of CMR in 2010 as covariates $\boldsymbol{X} = \{X_1, X_2, \cdots, X_7\}$, binarize the $PM_{2.5}$ in 2010 as treatment variable $T = 1(PM_{2.5} > 6)$, and study the CMR outcome Y in 2010.

5.2. Evaluation Process and Metrics

To assess the generalizability of our StableCFR on different subgroups and ensure reliable results, we present two types of data for testing our model: in-distribution (ID) data, which is drawn from the training distribution, and out-of-distribution (OOD) data, which is created by uniformly sampling from entire support. For both synthetic and Semi-PM-CMR datasets, we conduct 10 independent replications to report the mean and standard deviation of the estimation error over both ID and OOD datasets.

Over ID data and OOD data, we report the mean square errors (MSE) of the potential control/treated outcome estimation, the precision estimation of heterogeneous effect (PEHE $= \sqrt{n^{-1}\sum_{i=1}^{n}\left((\hat{y}_i(1) - \hat{y}_i(0)) - (y_i(1) - y_i(0))\right)^2}$), and the estimation error of average treatment effect ($\epsilon_{ATE} = |\hat{ATE} - ATE|$). While common evaluation metrics are useful, they may not be accurate for assessing the model's performance over subpopulations, particularly when dealing with underrepresented groups. Therefore, to demonstrate the stability of our algorithm, we report two additional metrics: the maximum error (MAXE) of PEHE for each individual, and the percentage ($P_{\downarrow 0.3}$) of individuals with errors that fall below the threshold of 0.3.

5.3. Baselines

We compare the proposed algorithm (**StableCFR**) with two groups of methods. One group is imbalance regression methods: (1) **Reweight** (Ren et al., 2022) weights data points using inverse re-weighting; (2) **BMC** (Ren et al., 2022) adopts batch-based monte-carlo with the balanced MSE; (3) **GAI** (Ren et al., 2022) adopts GMM-based analytical integration with the balanced MSE; (4) **DIRNet** (Yang et al., 2021) proposes FDS and LDS to smooth imbalance data; (5) **IPWNet** weights data points using inverse propensity score weighing. Another group is representation balance methods (elaborated in Sec.2.1): (1) **CFRNet** (Shalit et al., 2017); (2) **SITE** (Yao et al., 2018); (3) **CFRISW** (Hassanpour & Greiner, 2019a) (4) **DR-CFR** (Hassanpour & Greiner, 2019b) (5) **DeR-CFR** (Wu et al., 2022). Besides, as an ablation study, we use term **VANILLA** to denote the same network architecture that does not include any technique for dealing with imbalanced data.

5.4. Main Results

We report the main results on Syn-γ and Semi-PM-CMR, and the analysis on Real-PM-CMR in this section. The hyper-parameters experiments, the time complexity analysis, and additional results are provided in Appendix B.

Main Results on Syn-γ. We report the performance of different methods on the Syn-0.8 dataset in Tab. 1. From the results, we have the following observations: (1) Without

Table 1. The Generalization Experiments for Heterogeneous Treatment Effect on Syn-0.8 dataset.

	ID data (training distribution)				**OOD data (uniform distribution)**					
Method	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MAXE	$P_{\downarrow 0.3}$
VANILLA	**0.006(0.002)**	**0.005(0.004)**	**0.089(0.020)**	**0.003(0.019)**	0.139(0.043)	0.069(0.036)	0.500(0.113)	0.047(0.056)	10.044(1.745)	84.8%
Reweight	0.026(0.007)	0.353(0.246)	0.609(0.170)	0.055(0.069)	0.197(0.071)	1.008(0.319)	1.205(0.140)	0.155(0.114)	14.508(1.455)	31.4%
BMC	1.263(0.174)	2.246(0.322)	1.782(0.102)	0.499(0.108)	1.565(0.252)	0.816(0.146)	1.574(0.109)	0.361(0.123)	10.128(1.225)	8.1%
GAI	0.315(0.028)	0.245(0.045)	0.919(0.045)	0.032(0.020)	1.140(0.203)	1.462(0.513)	2.136(0.208)	1.323(0.148)	6.329(0.917)	19.1%
DIRNet	0.025(0.003)	0.073(0.012)	0.294(0.023)	0.003(0.023)	0.188(0.051)	0.465(0.119)	0.860(0.105)	0.025(0.042)	12.698(1.322)	52.5%
IPWNet	0.007(0.002)	0.009(0.004)	0.105(0.018)	0.003(0.021)	0.140(0.053)	0.095(0.044)	0.546(0.093)	0.031(0.089)	10.220(1.561)	80.5%
CFRNet	2.517(1.653)	2.634(1.731)	2.370(1.482)	0.124(0.103)	10.955(7.129)	16.377(10.66)	5.990(3.702)	3.582(2.362)	28.930(16.81)	26.9%
SITE	3.491(1.769)	1.933(1.093)	2.741(0.999)	0.058(0.060)	19.208(8.076)	11.341(6.235)	7.008(2.078)	4.086(1.399)	36.735(5.464)	7.4%
CFRISW	0.953(1.337)	0.875(1.359)	1.070(1.142)	0.047(0.069)	3.950(5.892)	5.400(9.204)	2.560(2.996)	1.212(2.010)	16.136(12.82)	37.6%
DRCFR	0.077(0.024)	0.065(0.012)	0.249(0.027)	0.008(0.039)	0.446(0.121)	0.315(0.091)	0.683(0.146)	0.007(0.083)	8.306(2.345)	60.4%
DERCFR	0.021(0.013)	0.020(0.007)	0.192(0.037)	0.007(0.065)	0.200(0.076)	0.223(0.103)	0.634(0.114)	0.128(0.144)	7.628(1.908)	64.4%
StableCFR	0.008(0.003)	0.006(0.003)	0.099(0.017)	0.009(0.017)	**0.046(0.020)**	**0.037(0.017)**	**0.299(0.054)**	**0.004(0.042)**	**5.889(1.577)**	**90.8%**

Table 2. The Generalization Experiments for Heterogeneous Treatment Effect on Semi-PM-CMR dataset.

	ID data (training distribution)				**OOD data (uniform distribution)**					
Method	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MAXE	$P_{\downarrow 0.5}$
VANILLA	**0.013(0.008)**	0.017(0.012)	**0.131(0.036)**	0.001(0.024)	0.304(0.099)	0.228(0.070)	0.555(0.116)	0.086(0.081)	6.206(1.224)	79.4%
Reweight	5.097(0.989)	1.726(0.628)	2.983(0.210)	1.268(0.330)	12.306(1.864)	5.935(1.388)	5.045(0.512)	1.897(0.407)	20.266(1.617)	9.5%
BMC	5.120(0.605)	2.461(0.811)	3.266(0.198)	0.253(0.368)	14.430(3.411)	8.762(1.410)	5.648(0.395)	0.892(0.486)	23.778(1.922)	8.3%
GAI	6.279(1.017)	2.126(0.394)	3.421(0.207)	0.217(0.311)	15.494(2.352)	8.393(0.988)	5.736(0.374)	0.789(0.585)	24.901(2.860)	8.7%
DIRNet	6.697(0.510)	3.162(0.159)	3.407(0.092)	2.147(0.110)	12.576(0.656)	4.089(0.339)	4.427(0.196)	2.400(0.251)	15.690(1.430)	8.5%
IPWNet	0.015(0.007)	**0.014(0.007)**	0.132(0.023)	0.002(0.014)	0.295(0.062)	0.199(0.045)	0.548(0.065)	0.077(0.042)	6.354(0.817)	80.7%
CFRNet	0.023(0.004)	0.028(0.009)	0.151(0.027)	**0.001(0.031)**	0.341(0.081)	0.246(0.057)	0.557(0.114)	0.063(0.077)	5.361(1.049)	79.6%
SITE	2.219(0.506)	2.032(0.198)	0.820(0.564)	0.006(0.060)	3.978(1.264)	3.615(0.579)	1.765(0.795)	0.264(0.177)	8.531(1.781)	31.4%
CFRISW	1.370(0.518)	1.470(0.391)	0.542(0.136)	0.027(0.082)	2.908(0.705)	3.194(0.289)	1.335(0.147)	0.225(0.123)	8.019(1.416)	34.8%
DRCFR	0.072(0.018)	0.090(0.013)	0.345(0.037)	0.013(0.032)	0.848(0.168)	0.677(0.092)	0.971(0.066)	0.110(0.074)	7.005(0.288)	52.0%
DERCFR	0.069(0.023)	0.083(0.023)	0.354(0.042)	0.001(0.106)	0.816(0.094)	1.072(0.189)	1.121(0.128)	0.118(0.125)	7.145(0.858)	45.9%
StableCFR	0.021(0.004)	0.028(0.005)	0.139(0.011)	0.011(0.009)	**0.271(0.056)**	**0.196(0.033)**	**0.489(0.051)**	**0.052(0.086)**	**4.926(0.604)**	**81.3%**

Table 3. The Heterogeneous Treatment Effect Estimation of $PM_{2.5} > 6$ on CMR from StableCFR by Adjusting the Variable Values of Real-PM-CMR (The Description is in Appendix A).

Adjust	Unemploy	Income	Female	Vacant	Owner	Edu	Poverty
0.3	7.673	6.658	5.729	1.276	5.568	10.310	5.039
0.2	7.179	6.351	5.471	2.902	5.905	9.043	5.423
0.1	6.688	6.202	5.670	4.549	6.127	7.675	5.817
0.0	6.206	6.206	6.206	6.206	6.206	6.206	6.206
-0.1	5.675	6.223	7.037	7.931	6.168	4.644	6.686
-0.2	5.162	6.338	8.075	9.646	5.973	2.971	7.093
-0.3	4.663	6.511	9.288	11.316	5.603	1.195	7.407

any prior knowledge about distribution, the conventional imbalance regression methods (Reweight, BMC, GAI, and DIRNet) can not accurately estimate the HTE on ID data or OOD data even if they balance the label distribution, since they do not explicitly consider the underrepresentation populations. (2) Without any regularity constraints, many CFR-based methods (CFRNet, SITE, and CFR-ISW) will overbalance and may map all input data to a constant vector in the face of severe underrepresentation distributions (we found this in our experiments). The disentanglement representation methods (DR-CFR and DeR-CFR) can alleviate the overbalance problem, but the decomposed representation still suffers from underrepresentation issues. (3) As ablation studies, we remove all techniques and found that VANILLA achieve the best HTE and ATE performance on the in-distribution (ID) data. However, as VANILLA did not take underrepresentation issues into account, its performance on the out-of-distribution (OOD) data was significantly worse than that of our StableCFR method. (4) Leveraging complex models to fit re-sampled data effectively, we improve the accuracy of HTE estimation on OOD data. Compared with the SOTA model, our StableCFR reduces the PEHE, ϵ_{ATE}, MAXE metrics by 40%, 43%, 7%, and improves 7% in $P_{\downarrow 0.3}$ metric. Our StableCFR aims to strike a balance between fitting the available data and generalizing well to out-of-distribution data. In the Syn-0.8 dataset, our StabkeCFR only slightly increases the PEHE on ID data compared to the best method (VANILLA), from 0.089 to 0.099. However, on OOD data, the PEHE is decreased from 0.500 to 0.299. These results demonstrate the effectiveness of our approach in improving performance in underrepresented subpopulations while maintaining good performance in the dominant subpopulation.

We extend our experiments to various underrepresentation scenarios and adjust the underrepresentation level in the Syn-γ datasets by varying the parameter $\gamma = \in \{0.5, 0.8, 1.0\}$. As shown in Figure 3, we observe that as the underrepresentation level decreased (from top to bottom), the generalization performance of CFRNet, VANILLA, and StableCFR improved. Overall, StableCFR demonstrates robust performance and provides reliable HTE estimation in most populations.

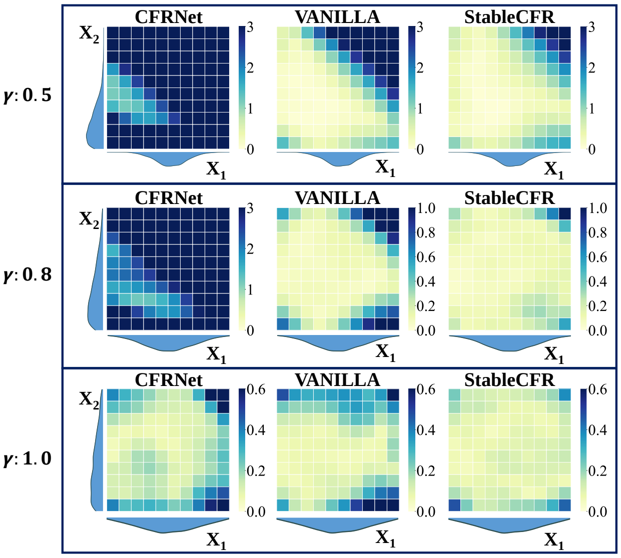


Figure 3. The PEHE Results (the lower the better) on Syn-γ Datasets for Adjusting the Underrepresentation Level by Changing the Parameter ($\gamma = 0.5, 0.8, 1.0$) of $\mathcal{N}(0, \gamma)$. We partition X_1 and X_2 into 10 equal parts according to the range of values, and then compute the local PEHE of the 10×10 region, in turn, to test the performance of the StableCFR under different OOD data.

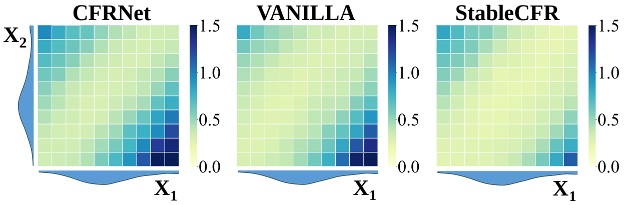


Figure 4. The PEHE Results on Semi-PM-CMR Dataset.

Main Results on Semi-PM-CMR.

To evaluate the effectiveness of our StableCFR on real-world applications, we conduct simulations using Semi-synthetic data with the real covariates of PM-CMR. As shown in Tab. 2, the performance of VANILLA, IPWNet, CFRNet, and StableCFR is similar in ID data due to the low underrepresentation level of the real variables. The difference between our StableCFR and the state-of-the-art methods is within 0.015 in each metric. On OOD data, our StableCFR outperforms the state-of-the-art model, reducing the PEHE, ϵ_{ATE}, MAXE metrics by 11%, 32%, and 8%, respectively, and improving by 1% in the $P_{\downarrow 0.3}$ metric. Additionally, in different 10×10 OOD subpopulations in Fig.4, our StableCFR maintains the most robust performance and achieves low HTE errors in most subpopulations.

Further Analysis on Real-PM-CMR. Since real-world datasets lack counterfactuals, as shown in Tab. 3, we implement the StableCFR algorithm by individually shifting the distribution of each variable in all cities to the left or right based on relative values of $\{-0.3, -0.2, -0.1, 0.0, 0.1, 0.2, 0.3\}$, while keeping the distributions of the other six variables fixed. Ranking of the fluctuation of HTE with the change of variables: Vacant(10.04) $>$ Edu(9.115) $>$ Female(3.817) $>$ Unempoly(3.01) $>$ Poverty(2.368) $>$ Owner(0.638) $>$ Income(0.456). The results demonstrate that counties with high levels of education and development are more susceptible to contracting CMR due to poor air quality. This is consistent with recent research findings that individuals working in high-tech industries may be more prone to neglecting physical exercise, as they often spend prolonged periods sitting in front of a computer screen. These regions may need to increase their focus on air quality and implement relevant policies. This is a demonstration of how our algorithm is connected to real-world applications.

6. Conclusion

In real applications, observational datasets are often constructed by pooling data from multiple sources or time periods, which can lead to concerns about representation issues. This problem is particularly pronounced when heterogeneous treatment effects are of primary interest, such as in precision medicine or targeted marketing. In the estimation of Heterogeneous Treatment Effects, conventional causal models have effectively addressed confounding bias, but they might unreliably estimate treatment effects and allocate suboptimal treatments for underrepresented populations. Our StableCFR can be seen as a robust HTE estimation scheme that improves the accuracy of HTE estimation for underrepresented populations. By filling the gap in techniques for stable HTE estimation and underrepresented issues, our StableCFR can provide a reliable and stable estimation of Heterogeneous Treatment Effects.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (62006207, 62037001, U20A20387), Young Elite Scientists Sponsorship Program by CAST (2021QNRC001), Zhejiang Province Natural Science Foundation (LQ21F020020), Project by Shanghai AI Laboratory (P22KS00111), Program of Zhejiang Province Science and Technology (2022C01044), the StarryNight Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-0010), and the Fundamental Research Funds for the Central Universities (226-2022-00142, 226-2022-00051). Bo Li's research was supported by the National Natural Science Foundation of China (No.72171131, 72133002); the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.

References

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

Abadie, A. and Imbens, G. W. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

Angrist, J. D. and Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Athey, S., Imbens, G. W., and Wager, S. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4): 597–623, 2018.

Branco, P., Torgo, L., and Ribeiro, R. P. Smogn: a preprocessing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pp. 36–50. PMLR, 2017.

Branco, P., Torgo, L., and Ribeiro, R. P. Rebagg: Resampled bagging for imbalanced regression. In *Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pp. 67–81. PMLR, 2018.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Erba, V., Gherardi, M., and Rotondo, P. Intrinsic dimension estimation for locally undersampled data. *Scientific reports*, 9(1):1–9, 2019.

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019a.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019b.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

He, Y.-Y., Wu, J., and Wei, X.-S. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 235–244, 2021.

Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., and Chang, B. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6626–6636, 2021.

Imai, K., Keele, L., and Yamamoto, T. Identification, inference and sensitivity analysis for causal mediation effects. 2010.

Imbens, G. W., Rubin, D. B., et al. Causal inference for statistics, social, and biomedical sciences. *Cambridge Books*, 2015.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Li, S., Vlassis, N., Kawale, J., and Fu, Y. Matching via dimensionality reduction for estimation of treatment effects. In *IJCAI*, pp. 3768–3774, 2016.

Liu, J., Sun, Y., Han, C., Dou, Z., and Li, W. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2970–2979, 2020.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009a.

Pearl, J. *Causality*. Cambridge university press, 2009b.

Ren, J., Zhang, M., Yu, C., and Liu, Z. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7926–7935, 2022.

Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American statistical Association*, 82(398):387–394, 1987.

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

Saltelli, A. Sensitivity analysis for importance assessment. *Risk analysis*, 22(3):579–590, 2002.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Shen, L., Lin, Z., and Huang, Q. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pp. 467–482. Springer, 2016.

Steininger, M., Kobs, K., Davidson, P., Krause, A., and Hotho, A. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Tang, K., Tao, M., Qi, J., Liu, Z., and Zhang, H. Invariant feature learning for generalized long-tailed classification. In *European Conference on Computer Vision*, pp. 709–726. Springer, 2022.

Torgo, L., Ribeiro, R. P., Pfahringer, B., and Branco, P. Smote for regression. In *Portuguese conference on artificial intelligence*, pp. 378–389. Springer, 2013.

VanderWeele, T. J. and Ding, P. Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274, 2017.

Wu, A., Yuan, J., Kuang, K., Li, B., Wu, R., Zhu, Q., Zhuang, Y. T., and Wu, F. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Wyatt, L. H., Peterson, G. C. L., Wade, T. J., Neas, L. M., and Rappold, A. G. Annual pm2. 5 and cardiovascular mortality rate data: Trends modified by county socioeconomic status in 2,132 us counties. *Data in brief*, 30: 105–318, 2020.

Xiang, L., Ding, G., and Han, J. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pp. 247–263. Springer, 2020.

Yang, Y., Zha, K., Chen, Y., Wang, H., and Katabi, D. Delving into deep imbalanced regression. In *International Conference on Machine Learning*, pp. 11842–11851. PMLR, 2021.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Yi, X., Tang, K., Hua, X.-S., Lim, J.-H., and Zhang, H. Identifying hard noise in long-tailed sample distribution. In *European Conference on Computer Vision*, pp. 739–756. Springer, 2022.

Zubizarreta, J. R. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A. The PM-CMR Datasets.

PM-CMR[2] (Wyatt et al., 2020) study the impact of $PM_{2.5}$ partical level on the cardiovascular mortality rate (CMR) in 2132 counties in the US using the data provided by the National Studies on Air Pollution and Health. As a real application, we select 7 variables (unemployment, income, female householder, vacant housing, owner-occupied housing, educational attainment, and families below the poverty level) associated with cardiovascular disease in 2010 as covariates and use the $PM_{2.5}$ partical level as treatment variable. The corresponding description of each variable is detailed in Tab. 4. The data presented in Fig. 5 displays skewed distributions resulting from an underrepresented population, where certain features have significantly fewer observations, i.e., few-shot samples. Such imbalanced data would introduce additional estimation error into HTE estimation. As shown in Tab. 5, we split the data into two parts with the median of each variable and calculated the difference between $PM_2.5$ and CMR for the two parts separately. To some extent, the results presented in Tab. 5 suggest that there is a confounding effect from each imbalanced covariate. The feature distributions illustrated in Fig. 5, demonstrate underrepresentation issues in the data.

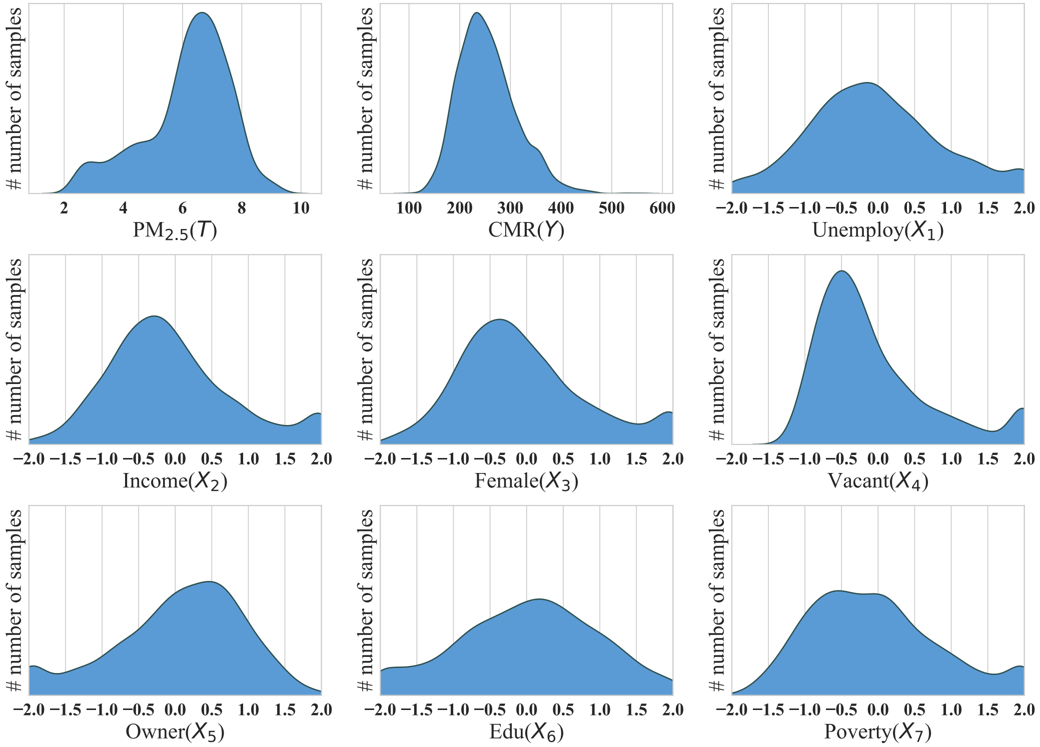


Figure 5. The Data Distribution on PM-CMR.

Table 4. The Description for Real Variables on Real-PM-CMR.

Variable	Description
$PM_{2.5}(T)$	Annual county PM2.5 concentration, $\mu\text{g}/\text{m}^3$
CMR(Y)	Annual county cardiovascular mortality rate, deaths/100,000 person-years
Unemploy(X_1)	Civilian labor force unemployment rate in 2010
Income(X_2)	Median household income in 2009
Female(X_3)	Family households - female householder, no spouse present in 2010 / Family households in 2010
Vacant(X_4)	Vacant housing units in 2010 / Total housing units in 2010
Owner(X_5)	Owner-occupied housing units - percent of total occupied housing units in 2010
Edu(X_6)	Educational attainment - persons 25 years and over - high school graduate (includes equivalency) in 2010
Poverty(X_7)	Families below poverty level in 2009

[2] PM-CMR:https://pasteur.epa.gov/uploads/10.23719/1506014/SES_PM25_CMR_data.zip

Table 5. The Confounding Effect from the Imbalanced Covariates.

	Unemploy	Income	Female	Vacant	Owner	Edu	Poverty
$PM_{2.5}$	0.302	0.208	0.765	-0.855	0.241	0.418	0.130
CMR	32.57	-51.70	45.78	22.12	1.491	35.78	54.34

Table 6. The Generalization Experiments for Heterogeneous Treatment Effect on Syn-0.5 dataset.

	ID data (training distribution)				**OOD data (uniform distribution)**					
Method	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MAXE	$P_{\downarrow 0.3}$
VANILLA	**0.003(0.001)**	**0.004(0.002)**	**0.073(0.012)**	**0.000(0.017)**	1.186(0.392)	3.575(1.366)	2.711(0.567)	0.923(0.377)	24.349(2.382)	48.6%
Reweight	0.026(0.016)	0.014(0.004)	0.190(0.046)	0.021(0.036)	1.813(0.585)	5.630(1.359)	3.350(0.429)	1.079(0.206)	27.393(1.661)	33.6%
BMC	2.192(0.304)	3.417(0.592)	2.199(0.102)	0.766(0.133)	4.014(0.973)	4.990(0.986)	3.223(0.303)	**0.362(0.449)**	23.087(1.941)	7.3%
GAI	1.115(0.129)	0.872(0.101)	1.507(0.026)	0.082(0.031)	15.808(2.808)	6.950(1.771)	6.151(0.589)	4.734(0.379)	**13.526(1.757)**	5.6%
DIRNet	0.013(0.008)	0.010(0.003)	0.125(0.031)	0.000(0.011)	1.464(0.380)	4.947(1.631)	2.886(0.517)	0.729(0.433)	24.900(1.956)	38.8%
IPWNet	0.003(0.001)	0.005(0.002)	0.073(0.012)	0.001(0.018)	0.978(0.339)	3.380(1.435)	2.497(0.600)	0.732(0.414)	22.589(3.169)	49.1%
CFRNet	0.964(0.051)	1.224(0.044)	1.576(0.038)	0.083(0.048)	18.868(0.206)	28.747(0.157)	9.370(0.029)	6.581(0.041)	41.385(0.041)	4.6%
SITE	0.211(0.022)	0.223(0.027)	0.714(0.049)	0.043(0.052)	12.585(0.899)	16.028(2.822)	7.239(0.297)	4.887(0.223)	37.374(1.118)	9.2%
CFRISW	0.480(0.400)	0.893(0.502)	1.123(0.527)	0.051(0.067)	9.870(7.575)	22.354(9.976)	6.878(2.667)	4.662(2.066)	33.626(9.101)	10.6%
DRCFR	0.021(0.008)	0.025(0.015)	0.176(0.040)	0.016(0.034)	1.196(0.270)	6.771(2.991)	3.009(0.652)	1.249(0.500)	23.023(2.965)	36.1%
DERCFR	0.009(0.004)	0.014(0.005)	0.145(0.019)	0.015(0.039)	1.178(0.399)	4.090(2.171)	2.610(0.513)	1.177(0.445)	18.330(3.183)	35.7%
StableCFR	0.005(0.001)	0.005(0.001)	0.085(0.011)	0.003(0.015)	**0.925(0.554)**	**1.722(0.849)**	**1.961(0.615)**	0.511(0.411)	18.407(3.363)	**54.4%**

B. Additional Experiments.

B.1. The Experiments for Adjusting the Underrepresentation Level

We expand experiments to various underrepresentation scenarios and adjust the underrepresentation level of Syn-γ datasets by changing the parameter $\gamma = 0.5, 0.8, 1.0$ in Tabs. 6, 7 & 8. From the results, we have the following observations:
(1) When the underrepresentation level is high ($\gamma = 0.5$ or 0.8 in Tabs. 6 & 7), the pure neural network will achieve SOTA performance on all metrics on ID data by fitting the training distribution without restriction. While the representation-based methods (CFRNet, SITE, and CFR-ISW) will overbalance and may map all input data to a constant vector in the face of severe underrepresentation distributions.
(2) In the experiments on the Syn-0.8 dataset (Tab. 7), StableCFR achieves SOTA performance on all metrics on OOD data. However, in the experiments on the Syn-0.5 dataset (Tab. 6), StableCFR does not perform as well as BMC and GAI methods on ϵ_{ATE} and MAXE metrics for OOD data, respectively. One possible reason is that BMC and GAI methods utilize balanced MSE as the training objective, which would reduce the maximum error by a penalty constraint term, but the error is still large on the Syn-0.8 dataset (Tab. 7).
(3) As the underrepresentation level decreases ($\gamma = 1.0$ in Tabs. 8), the estimation error from the neural network will decrease, and then the inverse propensity weighting method IPWNet will outperform the pure neural network method in the ID distribution, because the confounding bias caused by the imbalanced confounders will become the dominant error source.
(4) In Tabs. 7 & 8, through sampling from a multivariate uniform distribution and epsilon-greedy matching, our StableCFR balance confounders and smooth the imbalanced distribution to obtain the best HTE estimation on each individual (OOD data). Its gain is from increasing the HTE estimation performance on OOD data at the price of decreasing the estimation performance on ID data. Compared with the SOTA model, our StableCFR reduces the PEHE, ϵ_{ATE}, MAXE metrics by 40%, 43%, 7%, and improves 7% in $P_{\downarrow 0.3}$ metric in Tab. 7. Correspondingly, our StableCFR reduces the PEHE, ϵ_{ATE}, MAXE metrics by 15%, 40%, 21%, and improves 2% in $P_{\downarrow 0.3}$ metric in Tab. 8.

B.2. The Experiments for Adjusting the Hyper-parameters

In this paper, we split the samples on each dataset into training/validation data with an 80/20 proportion of training/validation splits. We return the best-evaluated iterate on validation data with early stopping and choose the best hyper-parameters from $\epsilon \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$[3] & $\sigma \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$.

The greedy strategy of searching for nearest neighbors for matching may lead to a frequent sampling of sparse samples in underrepresentation populations, hurting the model's generality. Therefore, we propose a hyper-parameters ϵ to trade-off exploration (distance-based sampling, with the receptive field controlled by hyper-parameter σ) and exploitation (nearest

[3] $\epsilon = 1.0$ denotes the greedy match algorithm without random exploration.

Table 7. The Generalization Experiments for Heterogeneous Treatment Effect on Syn-0.8 dataset.

	ID data (training distribution)				**OOD data (uniform distribution)**					
Method	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MAXE	$P_{\downarrow 0.3}$
VANILLA	**0.006(0.002)**	**0.005(0.004)**	**0.089(0.020)**	**0.003(0.019)**	0.139(0.043)	0.069(0.036)	0.500(0.113)	0.047(0.056)	10.044(1.745)	84.8%
Reweight	0.026(0.007)	0.353(0.246)	0.609(0.170)	0.055(0.069)	0.197(0.071)	1.008(0.319)	1.205(0.140)	0.155(0.114)	14.508(1.455)	31.4%
BMC	1.263(0.174)	2.246(0.322)	1.782(0.102)	0.499(0.108)	1.565(0.252)	0.816(0.146)	1.574(0.109)	0.361(0.123)	10.128(1.225)	8.1%
GAI	0.315(0.028)	0.245(0.045)	0.919(0.045)	0.032(0.020)	1.140(0.203)	1.462(0.513)	2.136(0.208)	1.323(0.148)	6.329(0.917)	19.1%
DIRNet	0.025(0.003)	0.073(0.012)	0.294(0.023)	0.003(0.023)	0.188(0.051)	0.465(0.119)	0.860(0.105)	0.025(0.042)	12.698(1.322)	52.5%
IPWNet	0.007(0.002)	0.009(0.004)	0.105(0.018)	0.003(0.021)	0.140(0.053)	0.095(0.044)	0.546(0.093)	0.031(0.089)	10.220(1.561)	80.5%
CFRNet	2.517(1.653)	2.634(1.731)	2.370(1.482)	0.124(0.103)	10.955(7.129)	16.377(10.666)	5.990(3.702)	3.582(2.362)	28.930(16.810)	26.9%
SITE	3.491(1.769)	1.933(1.093)	2.741(0.999)	0.058(0.060)	19.208(8.076)	11.341(6.235)	7.008(2.078)	4.086(1.399)	36.735(5.464)	7.4%
CFRISW	0.953(1.337)	0.875(1.359)	1.070(1.142)	0.047(0.069)	3.950(5.892)	5.400(9.204)	2.560(2.996)	1.212(2.010)	16.136(12.823)	37.6%
DRCFR	0.077(0.024)	0.065(0.012)	0.249(0.027)	0.008(0.039)	0.446(0.121)	0.315(0.091)	0.683(0.146)	0.007(0.083)	8.306(2.345)	60.4%
DERCFR	0.021(0.013)	0.020(0.007)	0.192(0.037)	0.007(0.065)	0.200(0.076)	0.223(0.103)	0.634(0.114)	0.128(0.144)	7.628(1.908)	64.4%
StableCFR	0.008(0.003)	0.006(0.003)	0.099(0.017)	0.009(0.017)	**0.046(0.020)**	**0.037(0.017)**	**0.299(0.054)**	**0.004(0.042)**	**5.889(1.577)**	**90.8%**

Table 8. The Generalization Experiments for Heterogeneous Treatment Effect on Syn-1.0 dataset.

	ID data (training distribution)				**OOD data (uniform distribution)**					
Method	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MSE(T=0)	MSE(T=1)	PEHE	ϵ_{ATE}	MAXE	$P_{\downarrow 0.3}$
VANILLA	0.007(0.002)	0.006(0.003)	0.093(0.013)	0.010(0.014)	0.051(0.037)	0.022(0.007)	0.247(0.070)	0.012(0.025)	5.244(1.858)	92.7%
Reweight	0.033(0.015)	0.237(0.127)	0.528(0.099)	0.020(0.069)	0.091(0.027)	0.495(0.152)	0.839(0.091)	0.109(0.082)	8.906(3.032)	41.3%
BMC	1.126(0.144)	1.457(0.191)	1.566(0.086)	0.417(0.058)	1.156(0.180)	0.634(0.073)	1.334(0.085)	0.124(0.051)	7.913(0.926)	11.2%
GAI	0.204(0.017)	0.144(0.015)	0.739(0.024)	0.001(0.043)	0.470(0.060)	0.447(0.066)	1.262(0.079)	0.629(0.060)	4.320(0.546)	27.4%
DIRNet	0.036(0.006)	0.099(0.013)	0.354(0.036)	0.014(0.009)	0.096(0.027)	0.278(0.058)	0.616(0.051)	0.045(0.028)	8.466(2.070)	55.1%
IPWNet	**0.006(0.002)**	**0.006(0.002)**	**0.087(0.009)**	**0.002(0.018)**	0.036(0.016)	0.021(0.006)	0.217(0.072)	0.010(0.029)	5.067(2.364)	93.8%
CFR	0.014(0.005)	0.018(0.007)	0.131(0.018)	0.012(0.034)	0.067(0.028)	0.056(0.032)	0.255(0.051)	0.005(0.049)	4.056(1.305)	88.6%
SITE	1.756(1.262)	2.089(1.181)	2.130(0.745)	0.251(0.120)	4.608(3.382)	4.437(4.277)	3.201(1.485)	1.375(0.779)	20.410(6.920)	14.4%
CFRISW	0.114(0.033)	0.119(0.025)	0.430(0.088)	0.011(0.050)	0.346(0.082)	0.363(0.126)	0.729(0.192)	0.017(0.082)	7.867(2.499)	52.5%
DRCFR	0.120(0.028)	0.101(0.027)	0.296(0.042)	0.021(0.021)	0.346(0.086)	0.235(0.045)	0.510(0.121)	0.035(0.063)	6.497(1.707)	64.5%
DERCFR	0.016(0.008)	0.021(0.009)	0.183(0.033)	0.024(0.052)	0.076(0.041)	0.083(0.041)	0.382(0.090)	0.074(0.079)	5.540(1.962)	76.0%
StableCFR	0.007(0.001)	0.007(0.001)	0.107(0.012)	0.006(0.014)	**0.023(0.012)**	**0.014(0.002)**	**0.184(0.045)**	**0.006(0.022)**	**3.994(1.542)**	**95.4%**

neighbor sampling). As shown in Fig. 2(b), we use the probability density function $f(\cdot|\sigma)$ of the normal distribution $\mathcal{N}(\mu = 0, \sigma)$ as the relative sampling probability. The closer the sample in training is to the point $\boldsymbol{q_j}$, the higher the probability of the sample being sampled. In this paper, we set three hyper-parameters $\{K, \epsilon, \sigma\}$ to denote top-K nearest neighbors, probability $\epsilon \in [0, 1]$ of greedy matching, and the standard deviation $\sigma\}$ of the normal distribution.

As demonstrated in Tab. 1 and Fig. 6, the pure uniform sampling with nearest neighbor sample ($\epsilon = 1$) has outperformed the best baseline in the PEHE on uniform data. Furthermore, based on the best-evaluated iterate, the uniform sampling with epsilon-greedy match ($K = 10, \epsilon = 0.6, \sigma = 0.25$) improves HTE estimation by avoiding frequent sampling of sparse samples. The experiments (Fig. 6) for adjusting the hyper-parameters demonstrate that the using parameters $\{K = 10, \epsilon = 0.6, \sigma = 0.25\}$ results in more stable HTE estimation with lower PEHE.

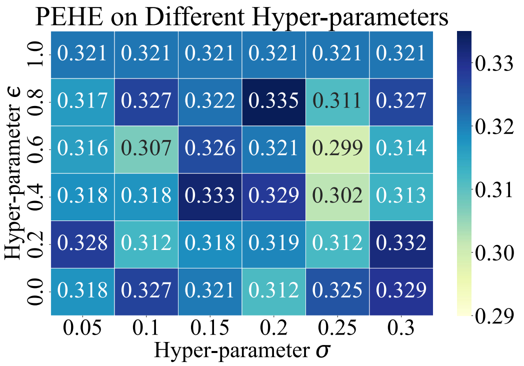


Figure 6. The Experiments for Adjusting the Hyper-parameters.

Table 9. Average Running Time(s) for Various Methods.

Method	Syn-γ (3000 units with 3 covariate)	Semi-PM-CMR (2132 units with 7 covariate)
VANILLA	85.5s	76.9s
Reweight	109.6s	107.2s
BMC	210.8s	193.4s
GAI	178.4s	169.5s
DIRNet	792.9s	653.1s
IPWNet	95.4s	94.0s
CFR	132.9s	120.0s
SITE	285.1s	243.9s
CFRISW	179.2s	173.6s
DRCFR	216.3s	203.3s
DERCFR	297.7s	266.4s
StableCFR	289.7s	287.5s

B.3. The Time Cost of Baselines and Further Analysis for StableCFR

The above algorithms are trained based on the same network architecture, but different training techniques will increase model complexity and training cost. In the synthetic and semi-synthetic datasets, we implement 10 replications to study the average running time(s) for the proposed model in a single execution and compare it to baselines. From the results (Table 9), we have the following observations: (1) The training cost of the pure neural network is only 85.5s on Syn-γ and 76.9s on Semi-PM-CMR, while the training cost of the DIRNet method is the largest, taking more than 10 minutes (600 seconds). (2) Except for DIRNet, all methods (including our StableCFR) can be executed in a single run within 300s, either on Syn-γ or on Semi-PM-CMR. (3) The single execution time for our StableCFR is 289.7s on Syn-γ and 287.5s on Semi-PM-CMR, which is within the acceptable range (600s).

Through sampling from a multivariate uniform distribution and epsilon-greedy matching, our StableCFR balance confounders and smooths the imbalanced distribution to obtain the best HTE estimation on each individual. As model complexity increases, both the time complexity and the single execution time of the model increase. In this section, we reformulate our calculations in matrix form:
(1) Calculate the distance matrix $\boldsymbol{D} \in \mathcal{R}^{mn}$ (Computational Complexity: $O(mnd)$):

$$\boldsymbol{D} = \|\boldsymbol{Q} - \boldsymbol{X}\|_2^2, \tag{12}$$

where, $\boldsymbol{Q} \in \mathcal{R}^{md}$ denotes the random points' matrix and $\boldsymbol{X} \in \mathcal{R}^{nd}$ denotes the covariates' matrix.
(2) Sort the distance matrix by row (Computational Complexity: $O(mn \log n)$):

$$sort(\boldsymbol{D}, 1), \tag{13}$$

(3) Calculate the probability density $\boldsymbol{P} \in \mathcal{R}^{mK}$ based on the distance of top-K points $\boldsymbol{KD} \in \mathcal{R}^{mK}$ (Computational Complexity: $O(mK)$):

$$f(\boldsymbol{KD}|\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\boldsymbol{KD}}{\sigma})^2} \tag{14}$$

(4) Match and sample the nearest neighbors with relative sampling probability P (Computational Complexity: $O(mK)$).

In summary, our StableCFR requires an additional computational cost to calculate the distance between samples, and this method is more suitable for small data. The computational complexity of the Epsilon-Greedy Matching is $O(mn(\log n + d) + 2mK)$, and the single execution time of our StableCFR is less than 300 seconds on 3000 samples. We believe this is acceptable. Besides, for a large dataset, we can randomly split the large data into multiple sub-datasets and then randomly select one sub-dataset to create the uniformed nearest neighbor batch.

Hardware used: Ubuntu 16.04.5 LTS operating system with 2 * Intel Xeon E5-2678 v3 CPU, 384GB of RAM, and 4 * GeForce GTX 1080Ti GPU with 44GB of VRAM.

Software used: Python 3.7.15 with TensorFlow 1.15.0, Pytorch 1.7.1, NumPy 1.18.0, and MatplotLib 3.5.3.