A Data-Driven Approach to Finding K for K Nearest Neighbor Matching in Average Causal Effect Estimation

Tingting Xu[1], Yinghao Zhang[1], Jiuyong Li[2], Lin Liu[2], Ziqi Xu[2], Debo Cheng[2(✉)], and Zaiwen Feng[1,3,4(✉)]

[1] College of Informatics, Huazhong Agricultural University, Wuhan, China
Zaiwen.Feng@mail.hzau.edu.cn
[2] UniSA STEM, University of South Australia, Adelaide, Australia
Debo.Cheng@unisa.edu.au
[3] Macro Agricultural Research Institute, Huazhong Agricultural University, Wuhan, China
[4] Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, China

Abstract. In causal inference, a fundamental task is to estimate causal effects using observational data with confounding variables. K Nearest Neighbor Matching (K-NNM) is a commonly used method to address confounding bias. However, the traditional K-NNM method uses the same K value for all units, which may result in unacceptable performance in real-world applications. To address this issue, we propose a novel nearest-neighbor matching method called DK-NNM, which uses a data-driven approach to searching for the optimal K values for different units. DK-NNM first reconstructs a sparse coefficient matrix of all units via sparse representation learning for finding the optimal K value for each unit. Then, the joint propensity scores and prognostic scores are utilized to deal with high-dimensional covariates when performing K nearest-neighbor matching with the obtained K value for a unit. Extensive experiments are conducted on both semi-synthetic and real-world datasets, and the results demonstrate that the proposed DK-NNM method outperforms the state-of-the-art causal effect estimation methods in estimating average causal effects from observational data.

Keywords: Causal inference · Causal effects estimation · Matching methods · Confounding bias · K nearest neighbor method

1 Introduction

Most scientific research in fields like medicine, economics, and behavioral science aims to infer the causal effect of a treatment on an outcome of interest [1,2]. One of the key challenges in causal inference is how to eliminate confounding

T. Xu and Y. Zhang—These authors contributed equally to this work.

F. Zhang et al. (Eds.): WISE 2023, LNCS 14306, pp. 723–732, 2023.
https://doi.org/10.1007/978-981-99-7254-8_56

bias (bias caused by confounders) when estimating causal effects [3]. Confounding bias arises from unbalanced distributions in treatment and control groups among covariates. Randomized control trials (RCTs) are the most effective way to eliminate the bias [4]. However, due to limitations of time, cost, and ethical concerns, RCTs are often infeasible [2]. As a viable alternative, estimating causal effects from observational data has become increasingly popular [5].

Matching is a popular approach to eliminating confounding bias by balancing the distribution of covariates between the treatment and control groups in paired matching [6]. Some commonly used matching methods include nearest-neighbor matching, subclassification matching, and weighting-based matching [7].

K nearest-neighbor matching (K-NNM) is one of the most widely used matching methods [1]. However, it is challenging to determine the K value. Setting K too large can lead to underfitting and increased computational bias. Conversely, setting K too small can increase the number of false matches. Moreover, the traditional KNN method assigns a fixed K value to all units, it may lead to a biased estimation and impractical outcomes in real-world applications [8,9]. Thus, there is a need to develop a novel K-NNM method that can set an optimal K value for each unit for causal effect estimation from observational data.

In this paper, we propose a novel Data-driven K-Nearest Neighborhood Matching algorithm (DK-NNM) for causal effect estimation from observational data. We first reconstruct all units to obtain a sparse coefficient matrix by using sparse representation learning, and the matrix is used to obtain the optimal K value of each unit. We then perform K-NNM based on propensity and prognostic scores for reducing the dimension of the covariates to avoid the curse of dimensionality [10]. The main contributions of this paper are summarized as follows:

- We utilize a sparse representation learning method to reconstruct the covariates for obtaining an optimal K value of each unit for unbiased estimation of causal effects with observational data.
- We utilize propensity and prognostic scores to reduce the dimension of covariates, instead of using all covariates or only propensity scores for matching.
- We apply our proposed DK-NNM method to both semi-synthetic and real-world datasets for estimating causal effects. The results demonstrate that our proposed DK-NNM method has better performance and efficiency compared with the state-of-the-art causal effect estimation methods.

2 Related Work

In the following, we review previous papers related to our proposed method. In practice, matching methods are commonly used in causal inference to identify groups with comparable or balanced covariate distributions [7]. Rubin and Thomas [11] projected the entire set of variables into one dimension and proposed propensity score matching. Diamond and Sekhon [12] proposed Genetic matching (GenMatch) to improve covariate balance by learning covariate weights, which can be thought of as a broader version of propensity score matching and Mahalanobis distance matching. Additionally, Leacy and Stuart [13] demonstrated

that using the combination of propensity and prognostic scores in matching methods had better performance than single-score-based matching methods in low-dimensional settings.

The most relevant work to ours is K-NNM. The exact K-NNM [1] is one of the most common methods. Luna et al. [14] adopted two resampling schemes to provide valid inference for K-NNM estimators. Additionally, Wager et al. [15] developed a random forest-based method to determine the weights for neighbor observations. However, these methods all use a fixed K value. When facing a more complicated scenario, it may lead to a huge deviation in the estimations.

3 Background

Under the potential outcome framework [2], we define the binary treatment variable T_i, which indicates whether a unit takes the treatment or not. The units that receive the treatment ($T_i = 1$) are treated units and the other units ($T_i = 0$) are control units. Let $\mathbf{X}$ denote the set of pre-treatment variables, i.e. covariates of T_i, which are variables that remain unchanged in the treatment process. Y_i represent the observed outcome for unit i, and we use $Y_i(1)$ to represent the potential outcome for unit i in treated group, and $Y_i(0)$ is the potential outcome of i in control group. However, in practice, each unit can only be assigned to either the treated or control group, so either $Y_i(1)$ or $Y_i(0)$ is unobserved for unit i. This is the fundamental challenge in causal effect estimation.

One of the basic tasks in causal inference is to estimate the causal effect of treatment T on outcome Y from observational data. In this paper, we would like to estimate the Average Treatment Effect (ATE) and the Average Treatment effect on the Treated group (ATT), as defined in the following:

$$ATE = E[Y_i(1) - Y_i(0)] \tag{1}$$

$$ATT = E[Y_i(1) \mid T = 1] - E[Y_i(0) \mid T = 1] \tag{2}$$

The propensity score $e(\mathbf{X})$ is defined as the conditional probability of a unit receiving the treatment conditioned on the full set of covariates $\mathbf{X}$ [16], i.e., $e(\mathbf{X}) = P(T = 1 \mid \mathbf{X})$. Furthermore, the prognostic score $p(\mathbf{X})$ is defined as the predicted outcome under the control condition [17], reflecting baseline "risk", i.e., $p(\mathbf{X}) = E[Y \mid \mathbf{X}]$. We use both the propensity and prognostic scores to construct a 2-dimensional space for conducting the matching process. For estimating causal effects from observational data, the following assumptions are required.

Assumption 1 (Stable Unit Treatment Value Assumption *[2]*). *The stability hypothesis has two implications: it first implies that the potential outcomes in different units are independent of each other. Besides, for each unit, there are no different forms of treatment levels that result in different potential outcomes.*

Assumption 2 (Overlap *[2]*). *For the covariates* $\mathbf{X}$*, every unit has a non-zero probability of receiving treatment* 1 *or* 0*. Formally,* $0 < P(T = t \mid \mathbf{X}) < 1, t = 0, 1$.

Assumption 3 (Unconfoundedness *[18]*). *The distribution of treatments is independent of the potential outcome conditioning on the set of covariates* $\mathbf{X}$, *i.e.* $T \perp\!\!\!\perp (Y(0), Y(1)) \mid \mathbf{X}$.

4 The Proposed DK-NNM Method

In this section, we introduce our proposed DK-NNM (Data-driven K-NNM) method. Specifically, we first explain the principle of selecting the K value for each unit through sparse representation learning. Then we present the specific process of our proposed DK-NNM method.

4.1 Sparse Representation Learning for the Optimal K Values

We propose to use sparse representation learning via self-representation for reconstructing the space of $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n and d are the numbers of units and covariates, respectively.

We use a linear model to represent a sample (unit) x_j as $x_j = x_i z_i + \varepsilon_i$ where z_i and ε_i are the dictionary and error term of x_i, respectively. The self-representation makes the reconstruction error as small as possible [19,20], so as to obtain the sparse coefficient matrix $\mathbf{Z}$. We utilize the least squares loss function to reconstruct the self-representation learning process:

$$\min_{z} \sum_{i=1}^{n} (x_i z_i - x_j)^2 = \min_{\mathbf{Z}} \|\mathbf{XZ} - \mathbf{X}\|_F^2 \tag{3}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times n}$ denotes the reconstruction coefficient matrix and is utilized to capture the correlations within the samples. Based on (3), we have $\mathbf{Z} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}$. In real-world applications, $\mathbf{X}^T\mathbf{X}$ is not always invertible. The ℓ_2-norm regularization term is often added to avoid the invertible issue. Therefore, our loss function can be rewritten as follows:

$$\min_{\mathbf{Z}} \|\mathbf{XZ} - \mathbf{X}\|_F^2 + \mu \|\mathbf{Z}\|_2^2 \tag{4}$$

where $\|\mathbf{Z}\|_2^2$ is the ℓ_2-norm regularization term and μ is a tuning parameter. The optimal solution of the optimization problem in (4) can be represented in a close form, i.e., $\mathbf{Z} = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}$. However, the solution of $\mathbf{Z}$ does not have the sparsity. To obtain the optimal k value for each unit, we expect that each unit is represented by those units strongly correlated with it, and the coefficients of the units with weak correlations are compressed to zero. Therefore, ℓ_2-norm is replaced by ℓ_1-norm in our method, which has been proved to generate sparsity [21]. Hence, our objective function is rewritten as follows:

$$\min_{\mathbf{Z}} \|\mathbf{XZ} - \mathbf{X}\|_F^2 + \alpha \|\mathbf{Z}\|_1, \mathbf{Z} \geqslant 0 \tag{5}$$

where $\|\mathbf{Z}\|_1$ is the ℓ_1-norm regularization term and $\mathbf{Z} \geqslant 0$ means that each element of $\mathbf{Z}$ is non-negative. And α is a tuning parameter of ℓ_1-norm to control the sparsity of matrix $\mathbf{Z}$. The larger α is, the more sparse the resulting matrix.

Moreover, we consider a nonlinear dimensionality reduction method, Locality Preserving Projections (LPP) [22] to preserve the neighborhood structure of the data. The LPP regularization term is defined as $\varphi(\mathbf{Z}) = Tr(\mathbf{Z}^T\mathbf{X}^T\mathbf{L}\mathbf{X}\mathbf{Z})$, where $\mathbf{L} \in \mathbb{R}^{d\times d}$ is a Laplacian matrix, implying the correlative information between features. $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix and $\mathbf{S} \in \mathbb{R}^{d\times d}$ is a similarity matrix. Thus, our final objective function is defined as:

$$\min_{\mathbf{Z}} \frac{1}{2} \|\mathbf{X}\mathbf{Z} - \mathbf{X}\|_F^2 + \alpha \|\mathbf{Z}\|_1 + \beta\varphi(\mathbf{Z}), \mathbf{Z} \geqslant 0 \tag{6}$$

where β is a tuning parameter to adjust the structure of matrix $\mathbf{Z}$, and is used to balance the magnitude between $\varphi(\mathbf{Z})$ and $\|\mathbf{X}\mathbf{Z} - \mathbf{X}\|_F^2$.

After optimizing (6), the optimal solution $\mathbf{Z}^*$ can be obtained, which represents the weight matrix. The element z_{ij} of $\mathbf{Z}^*$ can be understood as the correlation between the ith sample and the jth sample. If $z_{ij} > 0$, the ith sample and the jth sample are positively correlated; if $z_{ij} < 0$, they are negatively correlated; and if $z_{ij} = 0$, they are independent. To predict, we only use those relevant samples, i.e., samples with nonzero coefficients, instead of using all samples. We take an example to illustrate the optimal K value for each sample. We assume that the optimal solution $\mathbf{Z}^* \in \mathbf{R}^{4\times4}$ is as follows:

$$\mathbf{Z}^* = \begin{pmatrix} 0.5 & 0 & 0 & 0.4 & 0.6 \\ 0 & 0.3 & 0.8 & 0 & 0 \\ 0 & 0 & 0.7 & 0 & 0.1 \\ 0.4 & 0.8 & 0 & 0.6 & 0 \\ 0.6 & 0 & 0.1 & 0 & 0.8 \end{pmatrix}$$

There are five samples, and assume that the first two samples are in the treated group ($T = 1$) and the last three samples are in the control group ($T = 0$). There are three nonzero elements in the first row of $\mathbf{Z}^*$, that is z_{11}, z_{14} and z_{15}. It means that the first treated sample is only related to the last two samples except itself, i.e., the fourth and fifth control samples. Then the corresponding optimal K value for the first sample is 2. Similarly, the corresponding optimal K value for the second sample is 1. In this way, the K value of each sample can be obtained. Therefore, our proposed sparse representation learning method takes the data distribution and prior knowledge into account to choose the optimal K value for each sample respectively.

4.2 Matching Based on the Propensity and Prognostic Scores

After learning the optimal K value for each unit, we can employ the K-NNM method to estimate the causal effect from observational data. The matched unit can be considered as the counterfactual outcome [23] for the unit.

In the matching algorithm, our DK-NNM algorithm employs the Mahalanobis distance to measure the distance between each pair of units. Additionally, we transform all covariates into two-dimensional covariates based on propensity score $e(\mathbf{X})$ and prognostic score $p(\mathbf{X})$, significantly reducing dimensionality compared to full matching on the complete covariates. Leacy et al. [13] suggested that

matching method conditioning on the propensity and prognostic scores might help to decrease bias and enhance the accuracy of estimations. Hence, these two scores are utilized as the distance metric for our proposed DK-NNM method.

Formally, for the units i and j with estimated propensity scores $\hat{e}_i, \hat{e}_j$ and prognostic scores $\hat{p}_i, \hat{p}_j$, the score-based Mahalanobis distance between i and j can be defined as follows:

$$d(i,j) = \left[\begin{pmatrix}\hat{e}_i\\ \hat{p}_i\end{pmatrix} - \begin{pmatrix}\hat{e}_j\\ \hat{p}_j\end{pmatrix}\right]^{\top} \Sigma^{-1} \left[\begin{pmatrix}\hat{e}_i\\ \hat{p}_i\end{pmatrix} - \begin{pmatrix}\hat{e}_j\\ \hat{p}_j\end{pmatrix}\right], \tag{7}$$

where Σ indicates the variance-covariance matrix of $(\hat{e}, \hat{p})^{\top}$. The propensity score $\hat{e}$ can be estimated using logistic regression of the covariates $\mathbf{X}$ relative to the treatment variable T. For estimating the prognostic score, we restrict our analysis to the control group and perform an ordinary least squares regression of the outcome variable Y on the covariates $\mathbf{X}$. After conducting the K-NNM, we have a set of K-NNs for unit i, denoted as $\mathcal{J}_K(i)$. Then, we have

$$\tilde{Y}_i = (2T_i - 1)\frac{1}{K_i} \sum_{j \in \mathcal{J}_K(i)} Y_j \tag{8}$$

where K_i is the optimal k value for the ith unit, and $\tilde{Y}_i$ is the imputed outcome for unit i that is regarded as the unobserved potential outcome in this work.

5 Experiments

In this section, we conduct experiments on four datasets to evaluate the performance of our method, including IHPD, Jobs, Cattaneo2 and RHC. IHDP [24] is a semi-synthetic dataset, whose ground truth is generated by the synthetic process. The other three real-world datasets have empirical causal effects in the literature. We estimate ATT on the Jobs dataset, since the empirical ATT of this dataset is known. For experiments on the other datasets, we estimate ATE. To evaluate the performance of our proposed method, we use the root mean square errors (RMSE), and standard deviations (SD) as the evaluation metrics.

To demonstrate the superiority of our method, we take the following estimators for comparison: **NNM** (Nearest-Neighbor matching [11]), **PSM** (Propensity Score Matching [16]), **GenMatch** (Genetic Matching [12]), **BART** (Bayesian Additive Regression Trees [24]), **CF** (Causal Forest [25]), **BCF** (Bayesian Causal Forest [26]), and **S-LASSO** (S-learner using LASSO Regression [27]).

5.1 Experiment on Semi-synthetic Dataset IHDP

The dataset IHDP is a randomized experiment data [24] from the Infant Health and Development Program (IHDP). This program provided low-birth-weight, premature infants with high-quality child care and home visiting service. The IHDP dataset contains 747 units. There are 25 pre-treatment variables related to the study such as birth weight, twin status, first born, and sex. Following the procedures of Hill [24], the ground truth ATE of the IHDP dataset is 4.03.

Table 1. Experimental results on IHDP. The best performance is highlighted.

method	ATE	RMSE	SD
NNM	4.1300	0.1533	0.1645
PSM	3.8352	0.1945	0.1141
GenMatch	4.0002	0.0295	0.1874
BART	4.0180	3.5461	0.1192
CF	3.9733	2.7776	0.1019
BCF	3.9724	0.9369	0.9358
S-LASSO	3.8887	0.6956	0.6816
DK-NNM	4.0504	**0.0252**	**0.0145**

Table 2. Experimental results on Jobs. The best performance is highlighted.

method	ATT	RMSE	SD
NNM	198.16	683.34	1280.70
PSM	638.04	800.78	703.33
GenMatch	1098.50	212.54	1407.50
BART	966.60	2133.15	1024.00
CF	335.54	389.40	869.70
BCF	659.47	603.52	559.85
S-LASSO	277.24	2081.98	1992.62
DK-NNM	555.07	**331.41**	**88.94**

The experimental results of all estimators are listed in Table 1. We observe that DK-NNM achieves the best performance for ATE estimation, in particular, the standard deviation is extremely small against others. This demonstrates the competitiveness of DK-NNM with other state-of-the-art estimators.

Table 3. Experimental results on Cattaneo2 and RHC.

dataset	Cattaneo2		RHC	
method	ATE	SD	ATE	SD
NNM	−276.47	30.10	0.1006	0.0266
PSM	−303.97	18.63	0.0481	0.0143
GenMatch	−250.64	30.50	0.0671	0.0273
BART	−273.20	24.87	0.0389	0.0255
CF	−224.64	26.50	0.0229	0.0234
BCF	−227.08	79.81	0.0298	0.0176
S-LASSO	−240.55	87.95	0.0490	0.0810
DK-NNM	**−227.96**	**17.53**	**0.0656**	**0.0035**

5.2 Experiments on Three Real-World Datasets

Jobs. We adopt the Lalonde experiment dataset [28] and the control group data from the Panel Study of Income Dynamics(PSID) [29]. The pre-treatment variables include Age, Years of Education, and Proceeds in 1974 and 1975. The treatment variable represents whether the unit attends job training or not. We use the results of Imai's study as an experimental benchmark, the average causal effect on the treated group (ATT) is \$886 with a standard error of \$488 [29].

Table 2 shows the results for estimating ATT on the Jobs dataset. From Table 2, we see that the estimated ATT is close to the empirical ground truth and has the smallest SD. The results of BCF and BART are also close to the empirical ATT, but their estimated SDs are large. This indicates that the estimate of DK-NNM is consistent with the credible result and has a smaller SD.

Cattaneo2. The Cattaneo2 [30] is commonly used to study the effect of maternal smoking on birth weight. The mother's smoking status is used as the treatment variable. A variety of covariates are included, such as the Mother's marital status and Whether to drink alcohol or not. In a former paper, Almond et al. [31] studied the effect of maternal smoking on birth weight in pregnancy and they found a strong negative effect of about 200 g to 250 g lighter than normal birth.

The experimental results of all estimators on the Cattaneo2 dataset are listed in Table 3. The estimated ATE by DK-NNM is -227.96 g, which falls within the range of credible results obtained by [31]. And only the ATEs estimated by DK-NNM and CF fall within the empirical estimation interval (-250 g, -200 g), which further illustrates the competitiveness of our proposed DK-NNM method.

Right Heart Catheterization. The RHC dataset [32] was obtained from an observational study, which concerned the efficacy of Right Heart Catheterization (RHC) in the initial treatment of critically ill patients. The treatment is whether a patient received an RHC within 24 h. The covariates contain some physiological and clinical indicators. The existing evidence suggests that the use of RHC can result in higher 180-day mortality compared to not using RHC.

The experimental results of all estimators on the RHC dataset are described in Table 3. We can conclude that the causal effects obtained by PSM, GenMatch, BART, CF, BCF and DK-NNM are consistent. This indicates that the application of RHC leads to higher mortality rates within 180 days compared to not applying RHC. The consistent estimate of our method with the previous findings in the literature demonstrates the practicability of our DK-NNM method.

6 Conclusion

In this work, the key aspect of our proposed DK-NNM algorithm is to reconstruct the samples to obtain a sparse correlation matrix, allowing us to identify the best K value for each unit in the K-NNM method. To the best of our knowledge, this is the first study to incorporate data-driven K values into the K-NNM method in causal inference, while also considering the neighborhood

structure of the data. Additionally, we employ dimensionality reduction techniques to improve the efficiency of our DK-NNM method. Experimental results on extensive datasets demonstrate that our method outperforms other matching methods, indicating its potential usefulness for causal effect estimation in various applications.

Acknowledgements. We wish to acknowledge the support from the Australian Research Council (under grant DP200101210). This research project was also supported in part by the Major Project of Hubei Hongshan Laboratory under Grant 2022HSZD031, and in part by the Innovation fund of Chinese Marine Defense Technology Innovation Center under Grant JJ-2021-722-04, and in part by the open funds of Hubei Three Gorges Laboratory, and in part by the Fundamental Research Funds for the Chinese Central Universities under Grant 2662023XXPY004, 2662022JC004, and in part by the open funds of the National Key Laboratory of Crop Genetic Improvement under Grant ZK202203, Huazhong Agricultural University, and in part by the Inner Mongolia Key Scientific and Technological Project under Grant 2021SZD0099.

References

1. Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. **66**, 688–701 (1974)
2. Imbens, G.W., Rubin, D.B.: Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, Cambridge (2015)
3. Cheng, D., Li, J., et al.: Data-driven causal effect estimation based on graphical causal modelling: a survey. arXiv preprint arXiv:2208.09590 (2022)
4. Deaton, A., Cartwright, N.: Understanding and misunderstanding randomized controlled trials. Soc. Sci. Med. **210**, 2–21 (2018)
5. Cheng, D., Li, J., et al.: Causal query in observational data with hidden variables. In: ECAI 2020, pp. 2551–2558. IOS Press (2020)
6. Stuart, E.A.: Matching methods for causal inference: a review and a look forward. Stat. Sci. Rev. J. Inst. Math. Stat. **25**(1), 1–21 (2010)
7. Stuart, E.A.: Matching methods for causal inference: a review and a look forward. Stat. Sci.: Rev. J. Inst. Math. Stat. **25**(1), 1–21 (2010)
8. Zhang, S., Li, X., et al.: Learning k for kNN classification. ACM Trans. Intell. Syst. Technol. (TIST) **8**(3), 1–19 (2017)
9. Wu, W., Parampalli, U., et al.: Privacy preserving k-nearest neighbor classification over encrypted database in outsourced cloud environments. World Wide Web **22**, 101–123 (2019)
10. Cheng, D., Li, J., et al.: Sufficient dimension reduction for average causal effect estimation. Data Min. Knowl. Disc. **36**(3), 1174–1196 (2022)
11. Rubin, D.B.: Matching to remove bias in observational studies. Biometrics **29**, 159–183 (1973)
12. Diamond, A., Sekhon, J.S.: Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies. Rev. Econ. Stat. **95**(3), 932–945 (2013)
13. Leacy, F.P., Stuart, E.A.: On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. Stat. Med. **33**(20), 3488–3508 (2014)

14. de Luna, X., Johansson, P., Sjöstedt-de Luna, S.: Bootstrap inference for k-nearest neighbour matching estimators (2010)
15. Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. J. Am. Stat. Assoc. **113**(523), 1228–1242 (2018)
16. Rosenbaum, P.R., Rubin, D.B.: The central role of the propensity score in observational studies for causal effects. Biometrika **70**(1), 41–55 (1983)
17. Aikens, R.C., Greaves, D., Baiocchi, M.: A pilot design for observational studies: using abundant data thoughtfully. Stat. Med. **39**(30), 4821–4840 (2020)
18. Ye, S.S., Chen, Y., Padilla, O.H.M.: 2D score based estimation of heterogeneous treatment effects. arXiv preprint arXiv:2110.02401 (2021)
19. Zhu, X., Suk, H.-I., Shen, D.: Matrix-similarity based loss function and feature selection for Alzheimer's disease diagnosis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3089–3096 (2014)
20. Zhang, S., Cheng, D., et al.: Supervised feature selection algorithm via discriminative ridge regression. World Wide Web **21**, 1545–1562 (2018)
21. Zhu, X., Li, X., et al.: Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Trans. Neural Netw. Learn. Syst. **28**(6), 1263–1275 (2016)
22. He, X., Niyogi, P.: Locality preserving projections. In: Advances in Neural Information Processing Systems, vol. 16 (2003)
23. Chen, X., Wang, S., et al.: Intrinsically motivated reinforcement learning based recommendation with counterfactual data augmentation. World Wide Web 1–22 (2023)
24. Hill, J.L.: Bayesian nonparametric modeling for causal inference. J. Comput. Graph. Stat. **20**(1), 217–240 (2011)
25. Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. Ann. Stat. **47**(2), 1148–1178 (2019)
26. Hahn, P.R., Murray, J.S., Carvalho, C.M.: Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal. **15**(3), 965–1056 (2020)
27. Nie, X., Wager, S.: Quasi-oracle estimation of heterogeneous treatment effects. Biometrika **108**(2), 299–319 (2021)
28. LaLonde, R.J.: Evaluating the econometric evaluations of training programs with experimental data. Am. Econ. Rev. **76**, 604–620 (1986)
29. Imai, K., Ratkovic, M.T.: Covariate balancing propensity score. J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) **76**, 243–263 (2014)
30. Ghosh, T., Ma, Y., De Luna, X.: Sufficient dimension reduction for feasible and robust estimation of average causal effect. Stat. Sin. **31**(2), 821 (2021)
31. Almond, D., Chay, K.Y., Lee, D.S.: The costs of low birth weight. Q. J. Econ. **120**(3), 1031–1083 (2005)
32. Connors, A.F., et al.: The effectiveness of right heart catheterization in the initial care of critically III patients. JAMA **276**(11), 889–897 (1996)