Data Mining and Knowledge Discovery (2025) 39:35
https://doi.org/10.1007/s10618-025-01107-5

Data-driven learning optimal K values for K-nearest neighbour matching in causal inference

Yinghao Zhang[1] · Tingting Xu[1] · Debo Cheng[2] · Jiuyong Li[2] · Lin Liu[2] · Ziqi Xu[3] · Zaiwen Feng[1]

Received: 29 October 2024 / Accepted: 28 April 2025 / Published online: 28 May 2025
© The Author(s) 2025

Abstract

Within the realm of causal inference, a pivotal task involves causal effect estimation from observational data when there exist confounding variables. The K-Nearest Neighbour Matching (K-NNM) method is widely applied to handle confounding bias, but its general application sets a uniform K value for all samples, which can lead to suboptimal results in practice. To overcome this limitation, this paper introduces a novel method for causal effect estimation called Dynamic K-Nearest Neighbour Matching (DK-NNM). The DK-NNM method employs a data-driven learning strategy to determine the optimal value of K for each sample. In practice, DK-NNM reconstructs a sparse coefficient matrix for all samples using sparse learning, while simultaneously learning a graph matrix to preserve local information and sample similarity. This approach helps identify the most suitable K-value for each sample. Additionally, DK-NNM utilizes joint propensity and prognostic scores to effectively mitigate confounding bias arising from high-dimensional covariates during the K-NNM process. Experiments performed on various synthetic, semi-synthetic, and real-world datasets conclusively demonstrate that DK-NNM surpasses baseline models in estimating causal effects from observational data and provides significant improvements over traditional methods.

Keywords Causal inference · Confounding bias · Sparse learning · K-nearest neighbour matching

Responsible editor: Joao Gama.

Yinghao Zhang and Tingting Xu contributed equally to this work.

Extended author information available on the last page of the article

1 Introduction

Causal inference is essential to understand data generation mechanisms in a variety of real-world domains, including economics (Keele 2015), government policy evaluation (Holland et al. 1985), and fairness (Xu et al. 2023). A key focus in causal inference is estimating the causal effect of a treatment on an outcome of interest. A central challenge in this process is to address confounding bias, which occurs when covariates influencing both treatment and outcome have different distributions between the treated and control groups (Stuart and Elizabeth 2010). Although randomized controlled trials (RCTs) are considered the gold standard for establishing causality (Deaton and Cartwright 2018), they are often impractical due to high costs, time demands, and ethical concerns (Imbens and Rubin 2015). Consequently, estimating causal effects from observational data has become a pragmatic alternative in many applications (Cheng et al. 2024).

Matching is a cornerstone strategy in the domain of causal effect estimation, with the aim of alleviating confounding bias (Stuart and Elizabeth 2010). By matching, researchers strive to create comparable treated and control groups, thereby balancing the distribution of confounding variables. Commonly employed matching methods include exact matching, where units with identical covariate values are matched; propensity score matching (PSM), which pairs units based on their estimated probability of receiving treatment; full matching, which optimally matches all units; genetic matching (GenMatch) (Diamond and Sekhon 2013), a flexible and robust method that uses evolutionary algorithms to optimize balance; and Mahalanobis distance matching, which matches units based on a multivariate measure of distance (Stuart 2010). In our work, we focus on studying one of the most popular methods, K-Nearest Neighbour Matching (K-NNM) (Rubin 1974).

K-NNM is a widely used technique in causal inference (Stuart 2010) that seeks to pair each treated subject with K control subjects who share the closest covariate values, thereby forming comparable groups of treated and control units. However, the number of nearest neighbors, or the K parameter, plays a critical role in determining the quality of K-NNM for causal effect estimation. Selecting the appropriate K is challenging: choosing too small a K makes the estimate sensitive to outliers, while selecting too large a K reduces the similarity among matched samples, thereby failing to adequately mitigate confounding bias. Traditional approaches often use a fixed K, which can lead to poor estimates in real-world applications by overlooking heterogeneity within different data subsets (Zhang and Li 2017; Wu and Parampalli 2019).

To illustrate the limitations of a fixed K in K-NNM, we provide an example as depicted in Fig. 1. Here, counterfactual outcomes are calculated for three samples. When $K = 2$, the nearest neighbor space (represented by dashed circles) for the matched samples is defined. However, the second sample from the treatment group actually has three nearest neighbors, while the third sample has only one nearest control sample. This example underscores the importance of allowing different samples to be matched with varying numbers of nearest neighbors. Consequently, there is a clear need for a method that dynamically determines the

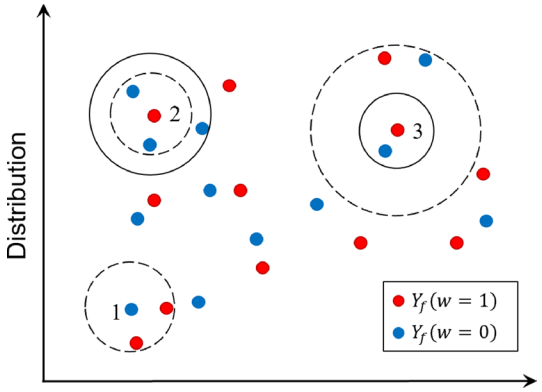


Fig. 1 An example is provided to illustrate the matching method and the limitations of a fixed K = 2 in K-NNM

optimal K-value, thereby enhancing K-NNM's accuracy and efficiency in causal effect estimation.

To determine an optimal K-value for each individual in K-NNM, we propose a sparse representation learning method that reconstructs a sparse coefficient matrix while simultaneously learning a graph matrix to preserve local information and sample similarity (Zhu and Li 2016). Consequently, the optimal K-value for each individualis derived from this learned sparse representation space. To the best of our knowledge, no prior work has explored the role of local structure information around samples in determining K-values in K-NNM for causal inference. Moreover, no established strategy exists for addressing confounding bias due to high-dimensional covariates when determining the optimal K-value for each individual. In our work, we employ both propensity and prognostic scores (Rubin and Thomas 2000; Leacy and Stuart 2014) to address confounding bias and reduce high-dimensional covariates, thereby circumventing the curse of dimensionality in matching process (Cheng and Li 2022). By jointly optimizing K values for each individual and making confounding adjustments, we develop a novel data-driven optimal K values for the K-NN matching method, referred to as DK-NNM.[1] Our primary contributions are outlined:

- We design a sparse learning-based method to reconstruct all samples and identify the optimal K value for each individual. To the best of our knowledge, our work is the first to concurrently learn sparse representations along with feature and sample correlations, enabling the determination of optimal K-values for each individual in K-NNM methods for causal inference.

[1] This manuscript is an expanded version of our recent conference paper, cited as Xu et al. (2023). In this version, we have significantly revised and enhanced the content to include in-depth discussions on the motivation and technical foundations of the DK-NNM method. Additionally, we provide a rigorous theoretical analysis of the sparse representation learning process, further substantiating our DK-NNM method. To comprehensively evaluate the effectiveness of DK-NNM in estimating causal effects, we have also incorporated new experiments, particularly those that use synthetic datasets, to complement the original analyses and offer a greater validation of the performance of the DK-NNM method.

- To mitigate confounding bias caused by high-dimensional covariates, we employ both prognostic and propensity scores to effectively reduce covariate dimensionality. Using these strategies, we propose a novel K-NN matching method, the DK-NNM method, for casual inference.
- The DK-NNM approach is evaluated using both synthetic datasets and real-world data. Experimental analysis demonstrates DK-NNM's effectiveness and efficiency, highlighting its notable advantage over existing methods for causal effect estimation.

2 Related work

There are two famous frameworks to address the confounding bias caused by covariates, i.e., the potential outcome framework (Rubin 1974) and the structural causal model (Pearl 1995). Our proposed method, DK-NNM, is conceived within the potential outcome framework. Subsequently, we conducted an in-depth review of prior literature relevant to our proposed DK-NNM method.

In practical applications, matching methods play a crucial role in causal inference by aiming to identify groups with comparable or balanced covariate distributions (Stuart 2010). The concept of optimal matching involves selecting matches by minimizing a global distance metric across all possible pairs (Gu and Rosenbaum 1993). Building on this concept, Rosenbaum (2017) introduced minimax and quantile constraints for dimensionality reduction. Rubin (1973) further contributed by proposing propensity score matching (PSM), which projects all covariates into a single dimension. Imbens (2004) refined this approach by adding regression adjustment. Diamond and Sekhon (2013) advanced these methods with GenMatch, which optimizes covariate balance by learning weights for covariates, building on both PSM and Mahalanobis distance matching. Additionally, Rubin and Thomas (2000) integrated propensity scores for prognostic covariates, underscoring the effectiveness of considering prognostic factors to reduce bias. Later simulation studies by Leacy and Stuart (2014) highlighted the advantages of combining propensity and prognostic scores to improve the quality of matching methods.

The most relevant work to our study is K-Nearest Neighbor Matching (K-NNM). The standard K-NNM method Rubin (1974) is widely used, with subsequent advancements enhancing its capability for estimating causal effects. Luna et al. (2010) proposed two resampling strategies to improve estimation accuracy in K-NN matching estimators. Wager and Athey (2018) introduced a tree-based K-NN approach using random forests to determine weights for neighboring observations, conceptualized as an adaptation of K-NN with an adaptive neighborhood metric. However, these methods uniformly use a fixed K value. When confronted with intricate scenarios, such as substantial differences between individuals, the adoption of a fixed K value may result in considerable deviations in causal effect estimation.

3 Background

We use the potential outcome framework as our basic model (Imbens and Rubin 2015). We consider the binary treatment variable T_i, where samples receiving treatment ($T_i = 1$) are referred to as treated samples, whereas those not receiving treatment ($T_i = 0$) are termed control samples. We use $\mathbf{X}$ to represent a set of pre-treatment covariates, which include $Pa(T)$ and $Pa(Y)$. This assumption ensures that $\mathbf{X}$ contains only relevant confounders, so there are no irrelevant noise variables. The observed outcome for sample i is denoted by Y_i. Here, $Y_i(1)$ and $Y_i(0)$ represent the potential outcomes for sample i if assigned to the treated group and the control group, respectively. Thus, the pair $(Y_i(1), Y_i(0))$ capture the potential outcomes for each sample. In real-life case, only one of $Y_i(1)$ and $Y_i(0)$ can be observed for an individual. This limitation presents the primary challenge in causal inference.

In the field of causal inference, a very important objective is to infer the impact of the treatment T on its outcome Y of interest using observational data. We aim to estimate the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated group (ATT). Their definitions are described as:

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

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

We also consider to use the propensity score, denoted as $e(\mathbf{X})$ Rosenbaum and Rubin (1983) and is defined as.

$$e(\mathbf{X}) = P(T = 1 \mid \mathbf{X}) \tag{3}$$

Moreover, the prognostic score, denoted as $p(\mathbf{X})$, has also been used in treatment effect estimation (Aikens et al. 2020). The prognostic score represents the baseline 'risk' associated with each individual and defined as follows.

$$p(\mathbf{X}) = E[Y \mid \mathbf{X}] \tag{4}$$

In causal inference, the following three assumptions are commonly made for ensuring that the causal effect can be estimated using observational data.

Assumption 1 (Stable individual Treatment Value Imbens and Rubin 2015) In data, it is assumed that the potential outcome in one individual is not affected by the specific treatment assignment in another individual. And there are no latent versions of the treatment leading to different potential outcomes for each individual.

Assumption 2 (Overlap Imbens and Rubin 2015) Each individual owns a non-zero probability to receive either treatment or control, given the covariates $\mathbf{X}$, i.e., $0 < P(T = t \mid \mathbf{X}) < 1$ for $t = 0, 1$.

Assumption 3 (Unconfoundedness Ye et al. 2021) The potential outcomes $(Y(0), Y(1))$ are conditionally independent of T given $\mathbf{X}$, i.e., $T \perp (Y(0), Y(1)) \mid \mathbf{X}$.

4 DK-NNM

In this section, we propose the DK-NNM method for causal effect estimation. DK-NNM first uses the local structure of $\mathbf{X}$ to learn a personalized K value for each individual, ensuring adaptive and flexible neighbor selection. Then, DK-NNM performs matching based on propensity scores and prognostic scores, which are estimated using the treatment (T) and outcome (Y), respectively. These two scores project the high-dimensional covariates into a low-dimensional space, effectively eliminating confounding bias introduced by high-dimensional covariates while also improving estimation accuracy by incorporating information from both the treatment and the outcome.

4.1 Determine the optimal K value

The idea of dynamically selecting the number of neighbors K in our matching method is conceptually similar to the strategy of adaptive bandwidth selection in kernel regression. In kernel regression, the smoothing parameter is dynamically adjusted based on local data density, allowing the bandwidth h for each sample, thus effectively balancing the bias-variance, which has been proved to be effective in several studies (Copeland 1997; Loader 1999). The DK-NNM method uses sparse learning to construct neighborhoods and adaptively selects matching samples K per sample based on the data neighborhood structure to optimize the quality of matching samples and reduce the bias in causal effect estimation. This connection provides a strong theoretical motivation for our proposed adaptive matching method.

The sparse representation learning through self-representation is proposed to reconstruct the space of $\mathbf{X} \in \mathbb{R}^{n\times d}$, where d and n stand for the numbers of covariates and samples, respectively.

The representation of the sample x_j in the linear model is expressed as $x_j = x_i z_i + \varepsilon_i$, where z_i represents the dictionary coefficients for the sample x_i, and ε_i is the error term associated with this representation. The goal of this self-representation is to minimize the reconstruction error, as detailed in prior studies (Zhu et al. 2014; Zhang and Cheng 2018). The primary objective is to derive a coefficient matrix $\mathbf{Z}$. To achieve this goal, we learn self-representation set $\mathbf{Z}$ by using the least squares loss function:

$$\min_{z} \sum_{i=1}^{n} (x_i z_i - x_j)^2 = \min_{\mathbf{Z}} \|\mathbf{XZ} - \mathbf{X}\|_F^2, \tag{5}$$

where $\mathbf{Z} \in \mathbb{R}^{n\times n}$ is the reconstructed representation matrix.

Expanding on Eq. (5), the expression for $\mathbf{Z}$ is derived as $\mathbf{Z} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}$. However, in practical scenarios, the matrix $\mathbf{X}^T\mathbf{X}$ may not be invertible. To circumvent this issue, we introduce an ℓ_2-norm regularization term, to mitigate the problem of invertibility. As a result, we can reformulate the loss function:

$$\min_{\mathbf{Z}} \|\mathbf{XZ} - \mathbf{X}\|_F^2 + \mu\|\mathbf{Z}\|_2^2, \tag{6}$$

where μ is a tuning parameter and $\|\mathbf{Z}\|_2^2$ represents the ℓ_2-norm regularization.

Equation (6) can be solved in a closed form as $\mathbf{Z} = (\mathbf{X}^T\mathbf{X} + \mu\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}$, where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is an identity matrix. Nevertheless, numerous studies have shown that this solution, $\mathbf{Z}$, lacks sparsity.

To identify the optimal K value for each sample, our goal is for each sample to be represented by those individuals exhibiting strong correlations with it. Additionally, we aim to compress the coefficients of individuals with weak correlations to zero. Hence, in our method, we use ℓ_1−norm to replace ℓ_2−norm, a modification proven to induce sparsity Zhu and Li (2016), and we have:

$$\min_{\mathbf{Z}} \|\mathbf{XZ} - \mathbf{X}\|_F^2 + \alpha\|\mathbf{Z}\|_1, \mathbf{Z} \geqslant 0, \tag{7}$$

where $\|\mathbf{Z}\|_1$ is the $\ell 1$−norm regularization, ensuring each value in $\mathbf{Z}$ remains non-negative, as indicated by $\mathbf{Z} \geqslant 0$. The parameter α acts as the tuning parameter for the ℓ_1−norm, playing a critical role in governing the sparsity level of $\mathbf{Z}$. A higher value of α leads to increased sparsity within the matrix.

To adapt our algorithm to complex high-dimensional data, we integrate a non-linear dimensionality reduction technique, i.e., Locality Preserving Projections (LPP) (He and Niyogi 2003). Unlike variance-based methods such as Principal Component Analysis (PCA), which emphasize global structure, LPP focuses on preserving local geometric relationships in the data. This ensures that samples with similar covariate patterns remain close in the transformed space, improving the stability of nearest-neighbor selection in high-dimensional settings.

LPP is formally defined as: $\varphi(\mathbf{Z}) = \mathrm{Tr}(\mathbf{Z}^T\mathbf{X}^T\mathbf{LXZ})$, where $\mathbf{L} \in \mathbb{R}^{d\times d}$ is the Laplacian matrix, capturing local feature similarities. The Laplacian matrix is constructed as: $\mathbf{L} = \mathbf{D} - \mathbf{S}$, with $\mathbf{S} \in \mathbb{R}^{d\times d}$ representing sample-wise similarities and D as the diagonal degree matrix. By embedding the data into a locally structured space, LPP enhances the robustness of the sparse representation step, reducing noise sensitivity and ensuring that neighbor selection remains reliable for causal inference.

Taking all these elements into account, our ultimate objective function is formulated as:

$$\min_{\mathbf{Z}} \frac{1}{2}\|\mathbf{XZ} - \mathbf{X}\|_F^2 + \alpha\|\mathbf{Z}\|_1 + \beta\varphi(\mathbf{Z}), \mathbf{Z} \geqslant 0, \tag{8}$$

where the tuning parameter β is to balance between $\varphi(\mathbf{Z})$ and $\|\mathbf{XZ} - \mathbf{X}\|_F^2$.

To prevent underfitting due to excessive regularization, we adhere to the common practice in sparse representation learning by setting the parameters α and β within the empirical range of $10^{-3} \sim 10^{-6}$. This range has been extensively utilized in studies on sparse optimization and graph-based learning (Zhou et al. 2003). Research indicates that when regularization parameters fall within this interval, they effectively constrain noise while balancing the sparsity of neighborhood structures (Wright et al. 2010). Within this range, we employ a grid search combined with cross-validation to identify the optimal parameter values. The search set is defined as $\{\alpha, \beta\} \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$.

Upon the successful optimization of (8), we obtain the optimal solution $\mathbf{Z}^*$. Each element z_{ij} quantifies the relative contribution of sample j in reconstructing

sample i, thereby capturing the intrinsic correlation between observations. To ensure the consistency and plausibility of these relationships, we enforce constraints on $\mathbf{Z}$ within the optimization formulation. Non-negativity constraint: $\mathbf{Z} \geqslant 0$ ensures that each element $z_{ij} \in [0, 1]$. $\mathbf{Z}$ represents the correlation between the *i*th and *j*th samples. Specifically, a positive z_{ij} (i.e., $z_{ij} > 0$) implies a positive correlation, indicating that the two samples move in the same direction. Importantly, a zero value (i.e., $z_{ij} = 0$) signifies independence between the samples. Since $\mathbf{Z}^*$ is obtained from real data through heuristic optimization, and all elements in $\mathbf{Z}^*$ are compressed during the sparse regularization process, we do not constrain the diagonal elements to be 1, so the diagonal elements are also dynamically determined by the optimization process.

In the context of prediction, this model prioritizes relevance by utilizing only those samples that have non-zero coefficients in the matrix $\mathbf{Z}^*$, as opposed to considering all available samples. To provide a clearer understanding of how the optimal K value is determined for each sample, let's examine an example with the optimal solution $\mathbf{Z}^*$ assumed to be in the space of $\mathbf{R}^{5\times5}$:

$$\mathbf{Z}^* = \begin{pmatrix} 0.7 & 0 & 0 & 0.3 & 0.6 \\ 0 & 0.8 & 0 & 0.4 & 0 \\ 0 & 0 & 0.3 & 0 & 0.2 \\ 0.3 & 0.4 & 0 & 0.5 & 0 \\ 0.6 & 0 & 0.2 & 0 & 0.8 \end{pmatrix}$$

In the given example with five individuals, we assume that the first two individuals are in the treatment group ($T = 1$), and the remaining three individuals are part of the control group ($T = 0$). Analyzing the first row of the matrix $\mathbf{Z}^*$, we identify three non-zero elements: z_{11}, z_{14}, and z_{15}. This indicates that the first treated individual has a correlation exclusively with the fourth and fifth individuals, which are in the control group. Consequently, the best K value for the first individual is determined to be 2, based on these correlations.

Similarly, the best K value for the second individual can be identified in the same manner. For instance, if the second row of $\mathbf{Z}^*$ indicates a nonzero element with only one of the control group individuals, then the best K value for the second treated individual would be 1.

This example illustrates how the sparse representation learning of sparse $\mathbf{Z}^*$ enables the individualized selection of the optimal K value for each individual. This tailored method ensures more precise and potentially more effective matching for causal inference.

The learned K values provide insight into how the model adapts to the local data structure, ensuring personalized and context-aware matching. In the example above, the sparse coefficient matrix $\mathbf{Z}^*$ serves as a learned similarity measure, where each row identifies the most relevant matches for a given individual. Nonzero elements in $\mathbf{Z}^*$ indicate a strong correlation between two samples, and the count of these elements determines the optimal K for that individual.

A smaller K value suggests that the individual has only a few highly relevant matches in the control group, indicating that the local data distribution is dense and

homogeneous. Conversely, a larger K value suggests that the individual resides in a more heterogeneous region, requiring a broader set of neighbors to achieve stable causal effect estimation. By dynamically adjusting K, our method ensures that each treated unit is matched with the most relevant control samples while avoiding arbitrary parameter selection.

4.2 Solving our objective function

Note that our objective function at Eq. (8) is convex function, but non-smooth, so we employ an accelerated proximal gradient method to solve it. We first perform the following accelerated proximal gradient operations on the objective function (8):

$$f(\mathbf{Z}) = \frac{1}{2}\|\mathbf{XZ} - \mathbf{X}\|_F^2 + \beta\varphi(\mathbf{Z}) \tag{9}$$

$$\vartheta(\mathbf{Z}) = f(\mathbf{Z}) + \alpha\|\mathbf{Z}\|_1, \tag{10}$$

where the objective function (8) is convex and differentiable.

Therefore, we use the proximal gradient to optimize $\mathbf{Z}$, and we initially iteratively update $\mathbf{Z}$ as:

$$\mathbf{Z}(m+1) = \arg\min_{\mathbf{Z}} \mathbf{G}_{\eta(m)}(\mathbf{Z}, \mathbf{Z}(m)) \tag{11}$$

$$\begin{aligned}\mathbf{G}_{\eta(m)}(\mathbf{Z}, \mathbf{Z}(m)) =& f(\mathbf{Z}(m)) + \langle \nabla f(\mathbf{Z}(m)), \mathbf{Z} - \mathbf{Z}(m)\rangle \\ &+ \frac{\eta(m)}{2}\|\mathbf{Z} - \mathbf{Z}(m)\|_F^2 + \alpha\|\mathbf{Z}\|_1\end{aligned} \tag{12}$$

$$\nabla f(\mathbf{Z}(m)) = (\mathbf{XX}^{\mathrm{T}} + \beta\mathbf{XLX}^{\mathrm{T}})\mathbf{Z}(m) - \mathbf{X}\hat{\mathbf{X}}^{\mathrm{T}}, \tag{13}$$

where $\langle ., ;. \rangle$ denotes the inner product operator, $\eta(m)$ determines the step size at the m-th mation, and $\mathbf{Z}(m)$ represents the value of $\mathbf{Z}$ obtained at the m-th iteration.

By neglecting the terms in Eq. (11) that are independent of $\mathbf{Z}$, it can be rewritten as follows.

$$\begin{aligned}\mathbf{Z}(m+1) &= \pi_{\eta(m)}(\mathbf{Z}(m)) \\ &= \arg\min \frac{1}{2}\|\mathbf{Z} - \mathbf{U}(m)\|_2^2 + \frac{\alpha}{\eta(m)}\|\mathbf{Z}\|_1,\end{aligned} \tag{14}$$

where $\mathbf{U}(m) = \mathbf{Z}(m) - \frac{\nabla f(\mathbf{Z}(m))}{\eta(m)}$ and $\pi_{\eta(m)}(\mathbf{Z}(m))$ represent the Euclidean projection of $\mathbf{Z}(m)$ onto the convex set $\eta(m)$.

Since $\mathbf{Z}(m+1)$ in each row (i.e., $\mathbf{z}^i(m+1)$) can be separated, weight updates can be performed separately for each row, as shown below,

$$\mathbf{z}^i(m+1) = \arg\min_{\mathbf{z}^i} \frac{1}{2}\left\|\mathbf{z}^i - \mathbf{u}^i(m)\right\|_2^2 + \frac{\alpha}{\eta(m)}\left\|\mathbf{z}^i\right\|_1, \tag{15}$$

where $\mathbf{u}^i(m) = \mathbf{z}^i(m) - \frac{\nabla f(\mathbf{z}^i(m))}{\eta(m)}$ and $\mathbf{z}^i(m)$ are the i-th row of $\mathbf{U}(m)$ and $\mathbf{Z}(m)$, respectively.

According to Eq. (15), $\mathbf{z}^i(m+1)$ will get a closed-form solution,

$$\mathbf{z}^{i*} = \max\left\{\left|\mathbf{z}^i\right| - \alpha, 0\right\} \cdot \operatorname{sgn}(\mathbf{z}^i), \tag{16}$$

where $\operatorname{sgn}(\mathbf{z}^i)$ denotes the sign function.

Additionally, to facilitate the accelerated proximal gradient method in Eq. (16), an auxiliary variable $\mathbf{V}(t+1)$ is introduced,

$$\mathbf{W}(m+1) = \mathbf{Z}(m) + \frac{\delta(m)-1}{\delta(m+1)}(\mathbf{Z}(m+1) - \mathbf{Z}(m)), \tag{17}$$

where the coefficient $\delta(m+1)$ is typically set as $\delta(m+1) = \frac{1+\sqrt{1+4\delta(m)^2}}{2}$.

Finally, we present our pseudo-code in Algorithm 1 for optimizing Eq. (8) and its convergence theorem. Convergence Theorem 1 follows from the accelerated gradient descent method introduced by Nesterov in convex optimization (Nesterov 2004). This theorem guarantees the convergence of the optimization process, which means that during the optimization process, $\mathbf{Z}$ gradually approaches the true solution and can obtain the optimal solution, making it identifiable under given constraints.

Theorem 1 Let $\{\mathbf{Z}(m)\}$ is a sequence produced by Algorithm 1. For all $\forall m \geq 1$, we have the following inequality holding:

$$\vartheta(\mathbf{Z}(m)) - \vartheta(\mathbf{Z}^*) \leq \frac{2\gamma L\left\|\mathbf{Z}(1) - \mathbf{Z}^*\right\|_F^2}{(m+1)^2}, \tag{18}$$

where γ is a positive predefined constant, L is the gradient Lipschitz constant for the function $f(\mathbf{Z})$ in (9), and $\mathbf{Z}^* = \underset{\mathbf{Z}}{\operatorname{argmin}}\ \vartheta(\mathbf{Z})$.

The convergence of Theorem 1 illustrates that $O(\frac{1}{m^2})$ is the convergence rate due to this accelerated proximal gradient method, in which m is the number of iterations in Algorithm 1.

Input: Initial step size $\eta(0)$, parameter $\alpha = 1$, and scaling factor γ.
Output: Solution matrix $\mathbf{Z}$.

Initialize iteration counter $m = 1$.
Initialize $\mathbf{Z}(1)$ with a random diagonal matrix.
repeat
  while $L(\mathbf{Z}(m)) > G_{\eta(m-1)}(\pi_{\eta(m-1)}(\mathbf{Z}(m)), \mathbf{Z}(m))$ **do**
    Update step size: $\eta(m-1) = \gamma\eta(m-1)$.
  end while
  Set $\eta(m) = \eta(m-1)$.
  Compute $\mathbf{Z}(m+1) = \arg\min_{\mathbf{Z}} G_{\eta(m)}(\mathbf{Z}, \mathbf{W}(m))$.
  Update $\delta(m+1) = \frac{1+\sqrt{1+4\delta(m)^2}}{2}$.
  Update $\mathbf{W}(m+1) = \mathbf{Z}(m) + \frac{\delta(m)-1}{\delta(m+1)}(\mathbf{Z}(m+1) - \mathbf{Z}(m))$.
  Increment iteration counter: $m = m + 1$.
until convergence of Eq. (12).

4.3 Matching over two scores

Upon determining the optimal K value for each individual, the K-NNM approach is then employed to estimate causal effects from the data. The DK-NNM method employs matching techniques to identify control samples with covariate distributions akin to the treatment samples. The essence of this matching method lies in simulating an RCT, where the matched individual serves as the counterfactual individual of the given individual.

In our DK-NNM approach, we implement the Mahalanobis distance to assess the dissimilarity between pairs of samples in the study. Moreover, a pivotal step in our approach involves transforming all covariates into a two-dimensional space. This transformation is guided by two key metrics: the propensity score, denoted as $e(\mathbf{X})$, and the prognostic score, represented by $p(\mathbf{X})$. By doing so, we significantly reduce the dimensionality of our data, which is a more efficient approach compared to full matching using the entire set of covariates. The combination of propensity scores, which balance the distribution of covariates between the intervention and control groups, and prognostic scores, which reduce random variation in the outcome variable, optimizes both covariate balance and outcome predictability during the matching process. This results in more comprehensive control of confounders and improved accuracy in causal effect estimation (Leacy and Stuart 2014). Consequently, these two scores serve as the distance metric in our DK-NNM method.

Precisely, for samples i and j characterized by estimated propensity scores $\hat{e}_i, \hat{e}_j (\hat{e} = P(T = 1|x))$, and prognostic scores $\hat{p}_i, \hat{p}_j (\hat{p} = E[Y|x])$, the Mahalanobis distance, rooted in scores, between samples i and j is formally defined as follows:

$$d(i,j) = \left[\begin{pmatrix} \hat{e}_i \\ \hat{p}_i \end{pmatrix} - \begin{pmatrix} \hat{e}_j \\ \hat{p}_j \end{pmatrix}\right]^{\top} \Sigma^{-1} \left[\begin{pmatrix} \hat{e}_i \\ \hat{p}_i \end{pmatrix} - \begin{pmatrix} \hat{e}_j \\ \hat{p}_j \end{pmatrix}\right], \tag{19}$$

where Σ represents the variance-covariance matrix of $(\hat{e}, \hat{p})^{\top}$.

Finally, we apply the K-NNM to identify a set of K-nearest neighbours for each individual i, which is denoted as $\mathcal{J}_K(i)$. The identification of these neighbours is crucial for the matching process. The subsequent step involves:

$$\tilde{Y}_i = \left(2T_i - 1\right) \frac{1}{K_i} \sum_{j \in \mathcal{J}_K(i)} Y_j \tag{20}$$

where K_i denotes the optimally determined value of K for the ith sample, while $\tilde{Y}_i$ represents the imputed outcome for this sample. It is important to note that $\tilde{Y}_i$ is treated as the unobserved potential outcome within the context of this study, also known as a counterfactual outcome.

Specifically, for each sample i, the individual causal effect (ICE) is defined as the difference between the observed outcome and the imputed counterfactual outcome.

For a treated sample $T_i = 1$, ICE is calculated as:

$$ICE_i = Y_i(1) - \tilde{Y}_i(0) \tag{21}$$

For a control sample $T_i = 0$, ICE is calculated as:

$$ICE_i = \tilde{Y}_i(1) - Y_i(0) \tag{22}$$

The Average Treatment Effect (ATE) is then computed by averaging the individual causal effects across all samples:

$$ATE = \frac{1}{N} \sum_{i}^{N} ICE_i \tag{23}$$

The Average Treatment Effect on the Treated (ATT) is computed by averaging the individual causal effects for the treated samples only:

$$ATT = \frac{1}{N_T} \sum_{i:T_i=1}^{N} ICE_i \tag{24}$$

The N_T denotes the number of treated samples.

5 Experiments

We evaluate our DK-NNM method on synthetic, semi-synthetic, and three real-world datasets to evaluate our method's effectiveness in causal effect estimation.

5.1 Experiment setting

The benchmark datasets used in our study include IHDP, Jobs, Cattaneo2, and RHC. The IHDP dataset Hill (2011) is characterized by a ground truth generated through a synthetic process. The others - Jobs, Cattaneo2, and RHC - are real-world datasets with empirically documented causal effects in the literature. For the Jobs dataset, we aim to estimate the ATT as its empirical ATT is established. In contrast, for the other datasets, we aim to estimate the ATE.

To evaluate the efficacy of the proposed method, we employ several metrics: Root Mean Square Error (RMSE) calculated as RMSE $= \sqrt{\frac{1}{M}\sum_{i=1}^{M}(ATE_i - \hat{ATE}_i)^2}$ or RMSE $= \sqrt{\frac{1}{M}\sum_{i=1}^{M}(ATT_i - \hat{ATT}_i)^2}$, Standard Deviations (SD) calculated as RMSE $= \sqrt{\frac{1}{M}\sum_{i=1}^{M}(\hat{ATE}_i - \overline{\hat{ATE}_i})^2}$, where M=30 represents the number of experiments, and Estimation Bias given by $\left|(\hat{ATE} - ATE)/ATE\right| \times 100\%$. These metrics assess accuracy (RMSE), variability (SD), and bias (Estimation Bias) in our estimations, ensuring a comprehensive evaluation of the proposed method.

To illustrate the superiority of our proposed method, we conduct comparisons with the estimators listed below: **NNM** (Nearest-Neighbour matching based on Mahalanobis metric Rubin 1973), **PSM** (Propensity Score Matching with logistic regression Rosenbaum and Rubin 1983), **GenMatch** (Genetic Matching, a weighted matching method learning weights by evolutionary search Diamond and Sekhon 2013), **BART** (Bayesian Additive Regression Trees focus on precise modeling of the response surface using a nonparametric Hill 2011), **CF** (Causal Forest based on random forest regression for estimating causal effect Athey et al. 2019), **BCF** (Bayesian Causal Forest Hahn et al. 2020), and **S-LASSO** (S-learner using LASSO Regression Nie and Wager 2021).

5.2 Evaluation using synthetic datasets

We generate the synthetic datasets with different sample sizes, i.e., 500, 1k, 2k and 3k for our experiments. Follow Cheng et al. (2024), we generate latent confounders F, M with Bernoulli distribution. For the observed variables $X = \{x_1, x_2, \cdots, x_p\}$, it is generated from the latent confounder F with independent Gaussian distributions as:

$$x_1, x_2, \cdots, x_p \sim N(F, \lambda * F), \tag{25}$$

where λ is the coefficient.

Then, generating the binary treatment Z with the Bernoulli distribution as the following function:

$$Z \sim Bernoulli(n, 1/(1 + exp(1 + 0.25 * M + 0.25 * F))) \tag{26}$$

Finally, two types of potential outcomes Y are generated to test the robustness of the model under various types of data scenarios. The one is a linear function as $Y_{linear} = 3 + 5 * \mathbf{Z} + 3 * \mathbf{M} + 2 * \mathbf{F} + \epsilon$, the other one is a nonlinear function as $Y_{nonlinear} = 3 + 5 * \mathbf{Z} + \mathbf{M} + (3 * \mathbf{M} + 2 * \mathbf{F}) * \mathbf{F} + \epsilon$, where ϵ is an error term.

Based on the above data generation process, we can know the true ITE for each sample and ATE over the whole dataset. In our configuration, the true ATE is 5. We conducted independent experiments on 30 datasets generated under each condition. This approach allows for a robust comparison across varying conditions while maintaining fairness in the evaluation against the baseline results.

To verify the effect of the DK-NNM's different components, we add ablation studies to the experiments on the synthetic data. By gradually removing or changing certain factors, explain our methods more comprehensively and make the research more reliable. All experiments of the ablation study are based on DK-NNM, including DK-NN (DK-NN method using the sparse matrices to represent causal effects), KNNM using a fixed K-values with both propensity and prognostic scores, e-KNNM (DK-NNM usingpropensity score only), P-KNNM (DK-NNM using prognostic score only).

5.2.1 Evaluation on causal effect estimation

We assess the performance of the DK-NNM using estimation bias and standard deviation. Table 1 presents the outcomes for linear datasets, while Table 2 displays the results for nonlinear datasets.

Results. Based on the analysis of experimental data, the following insights emerge: (1) NNM and PSM methods have large estimation bias and standard deviation on different synthetic dataset types because these methods rely on the assumption of no confounding and cannot effectively remove confounding factors. (2) Our proposed DK-NNM method achieves the smallest estimation bias

Table 1 Estimation bias (standard deviation) across 30 separate runs on synthetic data using Y_{linear}. The best results are bold-faced

Method	Sample size			
	500	1K	2K	3K
NNM	17.79 (0.56)	16.77 (0.43)	18.81 (0.34)	14.77 (0.38)
PSM	24.61 (0.63)	19.00 (0.36)	22.13 (0.30)	19.28 (0.31)
GenMatch	10.58 (0.64)	8.01 (0.50)	8.41 (0.41)	7.13 (0.36)
BART	17.22 (0.83)	8.25 (0.46)	6.47 (0.42)	5.22 (0.29)
CF	8.40 (0.48)	7.55 (0.42)	6.29 (0.38)	5.23 (0.30)
BCF	8.01 (0.52)	7.23 (0.34)	6.37 (0.42)	5.28 (0.30)
S-LASSO	9.56 (0.54)	6.96 (0.34)	7.13 (0.26)	5.32 (0.27)
DK-NN	37.23 (0.06)	37.89 (0.04)	36.13 (0.03)	37.55 (0.03)
KNNM	11.81 (0.71)	5.67 (0.31)	6.61 (0.27)	5.23 (0.29)
e-KNNM	10.79 (0.69)	5.77 (0.35)	6.61 (0.25)	5.48 (0.28)
p-KNNM	10.75 (0.67)	5.56 (0.35)	7.11 (0.29)	5.22 (0.27)
DK-NNM	**5.01 (0.36)**	**4.87 (0.31)**	**5.84 (0.22)**	**4.86 (0.25)**

Table 2 Estimation bias (standard deviation) across 30 separate runs on synthetic data using $Y_{nonlinear}$

Method	Sample size			
	500	1K	2K	3K
NNM	31.81 (0.44)	27.23 (0.22)	22.55 (0.23)	22.30 (0.18)
PSM	29.75 (0.63)	22.49 (0.37)	21.81 (0.36)	24.65 (0.30)
GenMatch	5.39 (0.30)	4.03 (0.25)	3.58 (0.22)	3.14 (0.18)
BART	4.59 (0.20)	4.11 (0.16)	3.18 (0.10)	3.04 (0.08)
CF	6.25 (0.21)	5.77 (0.18)	5.32 (0.11)	4.19 (0.08)
BCF	4.89 (0.19)	4.37 (0.16)	3.52 (0.11)	3.07 (0.09)
S-LASSO	5.24 (0.24)	5.80 (0.17)	6.35 (0.14)	5.73 (0.12)
DK-NN	31.60 (0.08)	34.45 (0.03)	34.35 (0.04)	32.54 (0.03)
KNNM	6.10 (0.27)	6.61 (0.18)	7.93 (0.13)	6.78 (0.12)
e-KNNM	5.11 (0.24)	3.69 (0.12)	3.83 (0.10)	4.62 (0.14)
p-KNNM	5.61 (0.22)	4.01 (0.13)	4.77 (0.10)	6.15 (0.11)
DK-NNM	**2.99 (0.22)**	**1.76 (0.12)**	**2.00 (0.12)**	**2.83 (0.14)**

The best results are bold-faced

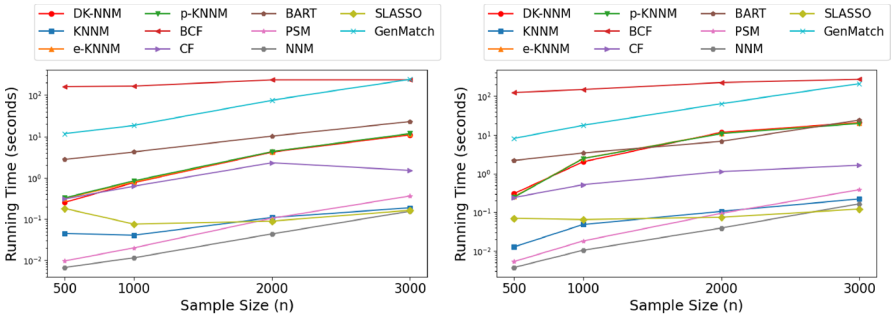
and standard deviation among all methods for both types of synthetic datasets, especially on nonlinear datasets. This indicates that our approach is adaptable to more complex data situations. (3) The experimental results of the BART, CF and BCF methods show that the larger the sample size, the smaller the estimation bias. However, our analysis, alongside previous research, shows that our method does not follow this trend. In large sample scenarios, k-nearest neighbor matching can become more sensitive to noise, potentially leading to overfitting issues. (4) In the ablation studies, the DK-NN method, which computes the causal effect solely from the K-nearest neighbors in a sparse matrix, exhibited a smaller standard deviation but a larger estimation bias due to the lack of debiasing with two scores. Although the other ablation methods yielded results comparable to those of the comparison methods, none matched the superior performance of the DK-NNM method

Consequently, the matched neighbours might not accurately reflect the true causal relationships within the actual distribution. Nevertheless, our method exhibits outstanding performance with small sample data, underscoring its significant contribution to real-world scenarios where sample sizes are limited.

5.2.2 Time cost of all methods

We have stored the average computation time for all methods when applied to synthetic datasets. These results are depicted in Fig. 2a and b, providing a visual comparison of the time efficiency across different methods.

Results. The BCF and GenMatch methods require particularly high processing times. In comparison, our proposed DK-NNM method demonstrates lower time consumption relative to the other methods.



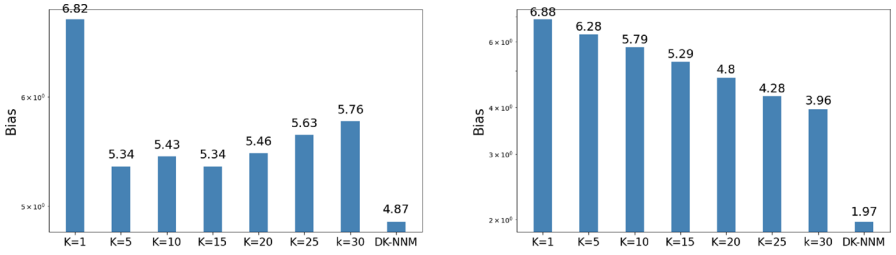
(a) With Y_{linear} (b) With $Y_{nonlinear}$

Fig. 2 The average cost time of all methods on the synthetic datasets in different sample sizes

5.2.3 The study of K values for KNNM against our DKNNM

To fully validate the effectiveness of our DKNNM method, we have designed a series of comparison experiments. The main purpose of the experiments is to test the performance advantage of our proposed method over the previous KNNM method with a fixed K-value for causal effect estimation. We use the previously generated linear and nonlinear datasets for these experiments, each with a sample size of 1000, to ensure both the breadth and reliability of our results. In our experiments, we set the K-value at various levels: 1, 5, 10, 15, 20, 25, and 30. For each K-value, we conducted 30 independent experiments to mitigate the impact of random errors on the data generation.

The experimental results are visualized in Fig. 3a and b. These results demonstrate that our proposed DKNNM method exhibits lower bias values for both linear and nonlinear datasets, significantly outperforming the traditional KNNM with a fixed k-value. This finding confirms the effectiveness of our method in reducing the bias of causal effect estimation across various data structures and complexities, thereby improving the accuracy of causal inference. Additionally, we found that the fixed k-value approach can lead to overfitting or underfitting in certain scenarios if



(a) With Y_{linear} (b) With $Y_{nonlinear}$

Fig. 3 Estimation bias on synthetic datasets with different K values

the K-value is not appropriately chosen, consequently increasing the bias in causal effect estimation. In contrast, our proposed method with an unfixed k-value avoids these issues by adapting more flexibly to the data's characteristics and automatically adjusting the number of matched neighbours.

5.3 Evaluation using semi-synthetic datasets

The IHDP dataset constitutes randomized experimental data Hill (2011) originating by the Infant Health and Development Program (IHDP). Commencing in 1985, the IHDP program conducted a trial providing intensive, high-quality childcare and home visiting services to low-birth-weight, premature infants. The IHDP dataset encompasses 747 samples, with 608 samples assigned to the control group and 139 samples to the treated group. It incorporates 25 pre-treatment covariates relevant to the program, e.g., firstborn status, sex, twin status, and maternal behavior during pregnancy and childbirth. Among these variables, 19 are binary covariates, and 6 are continuous covariates. Following Hill's procedures Hill (2011), the ground truth ATE for the IHDP dataset is determined to be 4.03.

We summarize all methods' results in Table 3. Additionally, we present the estimated causal effects and their corresponding 95% confidence intervals in Fig. 4. Based on Table 3, it's evident that our DK-NNM method outshines others in terms of accuracy in estimating ATE. especially with an extremely small standard deviation relative to other estimators. From Fig. 4, the estimations of GenMatch, BART, CF and DK-NNM are all close to the ground truth with a small standard deviation. This proves that our proposed DK-NNM is at least competitive with other methods on IHDP.

5.4 Evaluation using three real-world datasets

5.4.1 Jobs

Originating from a labor market intervention study, the Job training (Jobs) dataset merges the Lalonde experiment's dataset LaLonde (1986) with additional control data obtained from the Panel Study of Income Dynamics (PSID) Imai and

Table 3 Evaluation of causal effect estimations on IHDP, with the best-performing results 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**

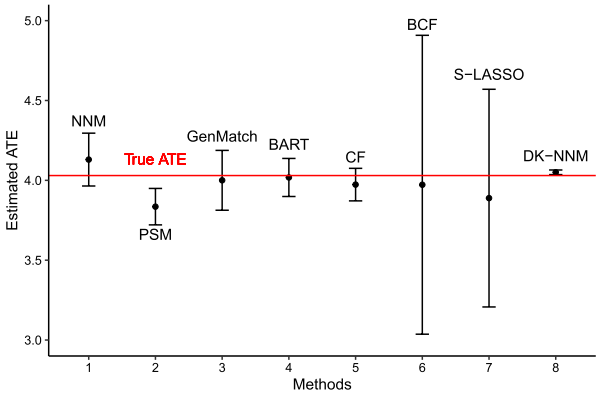


Fig. 4 Evaluation of causal effect estimations with a 95% confidence interval on the IHDP dataset. The empirical ATE is depicted by a red line

Ratkovic (2014). The pre-treatment covariates encompass Years of Education, Marital status, Ethnicity, Age, and Proceeds in 1974 and 1975. The treatment T is represented by an indicator variable indicating whether the sample participates in job training. The outcome of interest is the personal proceeds in 1978. The experimental benchmark for this dataset comes from Imai's study, with an ATT of $886 and a standard error of $488 Imai and Ratkovic (2014).

Table 4 presents the results of all methods for estimating ATT on the Jobs dataset, visually depicted in Fig. 5. Examination of Table 4 reveals that the estimated ATT closely aligns with the empirical ground truth, exhibiting the smallest SD and RMSE. Additionally, as observed in conjunction with Fig. 5, the results from the BCF, PSM, and BART methods also approximate the empirical ATT, though with higher estimated SDs. This indicates that the DK-NNM estimate is not only consistent with reliable results but also features a smaller SD.

Table 4 Evaluation of causal effect estimations on Jobs, with the best-performing results 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**

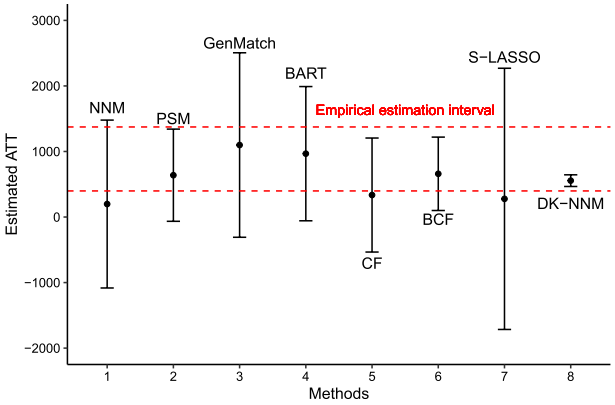


Fig. 5 Evaluation of causal effects with a 95% confidence interval on the Jobs dataset. The pair of dashed lines indicates the empirically estimated interval

5.4.2 Cattaneo2

The Cattaneo2 Ghosh et al. (2021), derived from a rich database of singletons in Pennsylvania, is frequently employed to investigate the impact of maternal smoking status on baby's birth weight. The mother's smoking status, i.e., smoking or non-smoking, served as the treatment variable. In this dataset, maternal smoking status (smoking or non-smoking) is the treatment variable. It includes 864 treated samples and 3,778 control samples, with birth weight as the outcome variable. The dataset also incorporates various covariates such as the mother's age, marital status, race, education, and alcohol consumption. A previous study by Almond et al. (2005) examined the effects of maternal smoking on birth weight, finding a significant negative impact, with infants weighing approximately 200 g to 250 g less than the average.

Table 5 presents all methods' results on Cattaneo2 and they are visualized in Fig. 6. DK-NNM yields an estimated Average Treatment Effect (ATE) of -227.96g,

Table 5 Evaluation of causal effects on the Cattaneo2 and RHC datasets, with the best-performing results highlighted

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**

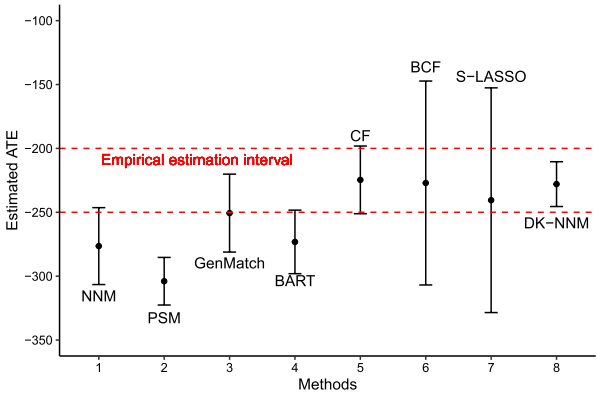


Fig. 6 Evaluation of causal effect estimations with 95% confidence intervals on the Cattaneo2 dataset. The pair of dashed lines indicates the empirically estimated interval (-250 g, -200 g)

aligning closely with the credible results reported by Almond et al. (2005). Figure 6 shows that only the estimated ATEs by DK-NNM and CF fall within the empirically estimated interval (-250 g, -200 g), emphasizing the competitiveness of the DK-NNM method.

5.4.3 RHC

The Right Heart Catheterization (RHC) dataset Connors et al. (1996) originates from an observational study examining the effectiveness of RHC as an initial treatment for critically ill patients. The treatment variable indicates whether patients underwent RHC within 24 h of admission, while the outcome variable records whether a patient died within 180 days following admission. The dataset includes various standard physiological and clinical covariates. Existing study shows that using RHC may be associated with higher 180-day mortality rates against not using RHC.

Results are compiled in Table 5 and visualized in Fig. 7. The causal effects derived from methods such as PSM, GenMatch, BART, CF, BCF, and DK-NNM demonstrate consistency. It suggests that the use of RHC is associated with increased mortality within 180 days compared with no use of RHC. The concordance of the estimates from our method with prior research underscores the DK-NNM's practicality and reliability.

6 Conclusion

In our research, the DK-NNM method focuses on reconstructing individuals to produce a sparse representation matrix. This key strategy is crucial for determining the optimal K value for each individual in the K-NNM method. Our study is the first to integrate a data-driven approach for discovering K values into K-NNM for

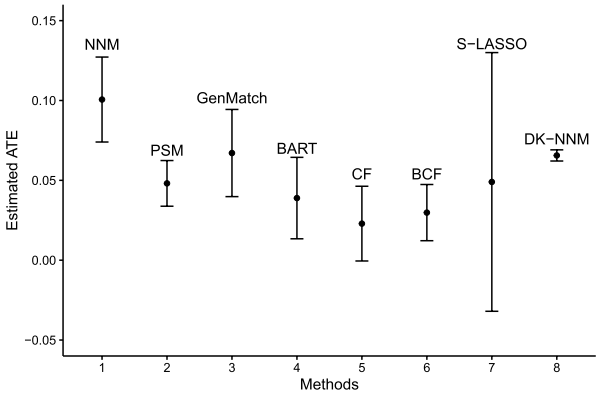


Fig. 7 Evaluation of causal effect estimations with a 95% confidence interval on the RHC dataset

estimating causal effects using observational data, while simultaneously considering the neighborhood structure inherent in the data. Our method addresses the issue of fixed K values in the traditional K-NNM method. In conventional approaches, the value of K is typically determined by the user based on prior experience. In contrast, the K value in our algorithm is learned and selected in a data-driven manner. As a result, our approach can be applied in scenarios where users are unable to choose an appropriate K value based on experience, significantly reducing the time required for testing different K values. We also apply two scores to reduce dimensionality and mitigate confounding bias, making our approach more suitable for complex datasets encountered in real-world applications. We conducted experiments on large datasets to highlight our method's superior performance compared to other matching techniques. These findings suggest the significant potential of DK-NNM in accurately estimating causal effects in various settings.

Future work could further explore the scalability and efficiency improvements of the algorithm. Although our method effectively reduces the computational burden through sparse constraints and accelerated approximate gradient methods, there is still room for improvement. For larger scale applications, data subsampling strategies and distributed computing techniques can be explored to improve algorithm performance. These scalability considerations ensure that our approach remains computationally feasible for large data sets while maintaining theoretical advantages.

Acknowledgements The authors would like to thank the Action Editors and the Reviewers for their valuable comments.

Author Contributions Yinghao Zhang: Writing - review & editing, Supervision, Funding acquisition, Conceptualization. Tingting Xu: Writing - review & editing, Writing - original draft, Visualization, Validation. Debo Cheng: Writing - review & editing, Formal analysis, Methodology, Conceptualization. Jiuyong Li: Writing - review & editing, Formal analysis, Investigation, Conceptualization. Lin Liu: Writing - review & editing, Methodology, Conceptualization. Ziqi Xu: Writing - review & editing, Methodology, Conceptualization. Zaiwen Feng: Writing - review & editing, Supervision, Funding acquisition, Conceptualization.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFF1000100, the Fundamental Research Funds for the Chinese Central Universities under Grant 2662023XXPY004, and the Australian Research Council under Grant DP230101122.

Data Availability The IHDP dataset is available on https://github.com/vdorie/npci. The Jobs dataset can be obtained from https://github.com/jjchern/lalonde. The Cattaneo2 dataset is available on http://www.stata-press.com/data/r13/cattaneo2.dta. The RHC dataset is available on https://cran.r-project.org/web/packages/Hmisc/index.html.

Declarations

Ethical approval Not applicable.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Aikens RC, Greaves D, Baiocchi M (2020) A pilot design for observational studies: using abundant data thoughtfully. Stat Med 39(30):4821–4840

Almond D, Chay KY, Lee DS (2005) The costs of low birth weight. Q J Econ 120(3):1031–1083

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. Ann Stat 47(2):1148–1178

Cheng D, Li J et al (2022) Sufficient dimension reduction for average causal effect estimation. Data Min Knowl Disc 36(3):1174–1196

Cheng D, Li J, Liu L, Liu J, Le TD (2024) Data-driven causal effect estimation based on graphical causal modelling: A survey. ACM Comput Surv 56(5):1–37

Connors AF, Speroff T, Dawson NV, Thomas C, Harrell FE, Wagner D, Desbiens N, Goldman L, Wu AW et al (1996) The effectiveness of right heart catheterization in the initial care of critically iii patients. JAMA 276(11):889–897

Copeland KA (1997) Local polynomial modelling and its applications. Taylor & Francis, New York

Deaton A, Cartwright N (2018) Understanding and misunderstanding randomized controlled trials. Soc Sci Med 210:2–21

Diamond A, Sekhon JS (2013) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Rev Econ Stat 95(3):932–945

Ghosh T, Ma Y, De Luna X (2021) Sufficient dimension reduction for feasible and robust estimation of average causal effect. Stat Sin 31(2):821

Gu XS, Rosenbaum PR (1993) Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat 2(4):405–420

Hahn PR, Murray JS, Carvalho CM (2020) Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). Bayesian Anal 15(3):965–1056

He X, Niyogi P (2003) Locality preserving projections. Adv Neural Inf Process Syst 16

Hill JL (2011) Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 20(1):217–240

Holland PW, Glymour C, Granger C (1985) Statistics and causal inference*. ETS Res Rep Ser 1985(2):72

Imai K, Ratkovic MT (2014) Covariate balancing propensity score. J R Stat Soc: Ser B (Stat Methodol) 76

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86(1):4–29
Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press, Cambridge
Keele L (2015) The statistics of causal inference: a view from political methodology. Polit Anal 23(3):313–335. https://doi.org/10.1093/pan/mpv007
LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev pp 604–620
Leacy FP, Stuart EA (2014) 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
Loader C (1999) Local regression and likelihood. Springer, New York
Luna X, Johansson P, Sjöstedt-de Luna S (2010) Bootstrap inference for k-nearest neighbour matching estimators
Nesterov Y (2004) Introductory lectures on convex optimization: a basic course, vol 87. Springer, New York
Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 108(2):299–319
Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–688
Rosenbaum PR (2017) Imposing minimax and quantile constraints on optimal matching in observational studies. J Comput Graph Stat 26(1):66–78
Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41–55
Rubin DB (1973) Matching to remove bias in observational studies. Biometrics, 159–183
Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701
Rubin DB, Thomas N (2000) Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc 95(450):573–585
Stuart EA (2010) Matching methods for causal inference: a review and a look forward. Stat Sci A Rev J Inst Math Stat 25(1):1–21
Stuart EA (2010) Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics 25(1):1–21
Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242
Wright J, Ma Y, Mairal J, Sapiro G, Huang TS, Yan S (2010) Sparse representation for computer vision and pattern recognition. Proc IEEE 98(6):1031–1044
Wu W, Parampalli U et al (2019) Privacy preserving k-nearest neighbor classification over encrypted database in outsourced cloud environments. World Wide Web 22:101–123
Xu Z, Liu J, Cheng D, Li J, Liu L, Wang K (2023) Disentangled representation with causal constraints for counterfactual fairness. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 471–482. Springer
Xu T, Zhang Y, Li J, Liu L, Xu Z, Cheng D, Feng Z (2023) A data-driven approach to finding k for k nearest neighbor matching in average causal effect estimation. In: International Conference on Web Information Systems Engineering (WISE), pp 723–732. Springer
Ye SS, Chen Y, Padilla OHM (2021) 2d score based estimation of heterogeneous treatment effects. arXiv preprint arXiv:2110.02401
Zhang S, Cheng D et al (2018) Supervised feature selection algorithm via discriminative ridge regression. World Wide Web 21:1545–1562
Zhang S, Li X et al (2017) Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST) 8(3):1–19
Zhou D, Bousquet O, Lal T, Weston J, Schölkopf B (2003) Learning with local and global consistency. Adv Neural Inf Process Syst 16
Zhu X, Li X et al (2016) Robust joint graph sparse coding for unsupervised spectral feature selection. IEEE Trans Neural Netw Learn Syst 28(6):1263–1275
Zhu X, Suk H-I, Shen D (2014) 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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yinghao Zhang[1] · Tingting Xu[1] · Debo Cheng[2] · Jiuyong Li[2] · Lin Liu[2] · Ziqi Xu[3] · Zaiwen Feng[1]

✉ Debo Cheng
chedy055@mymail.unisa.edu.au

✉ Zaiwen Feng
Zaiwen.Feng@mail.hzau.edu.cn

Yinghao Zhang
yhzhang@mail.hzau.edu.cn

Tingting Xu
xutingting@webmail.hzau.edu.cn

Jiuyong Li
Jiuyong.li@unisa.edu.au

Lin Liu
lin.liu@unisa.edu.au

Ziqi Xu
ziqi.xu@rmit.edu.au

[1] College of Informatics, Huazhong Agricultural University, Wuhan 430070, Hubei, China

[2] UniSA STEM, University of South Australia, Adelaide, SA 5095, Australia

[3] School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia