Toward efficient matching and weighting for estimating the causal effects of multiple continuous exposures

Yihui He, Xiao Wu

November 11, 2024

Abstract

Keywords: bias-corrected matching, continuous treatment, double machine learning, observational studies, semiparametric

1 Introduction

Climate change is and will remain the most important threat to human health and welfare. As emissions continue to drive temperatures higher, climate threats (e.g., wildfires, heat waves, hurricanes, etc.) increase in frequency and intensity. These changes in climate-relevant exposures are linked to numerous adverse health outcomes and will continue to negatively impact human health. Compound climate events, significant impacts resulting from the combination of interacting physical processes across multiple spatial and temporal scales, frequently give rise to floods, wildfires, heat waves, and droughts. To date, climate change and health analyses have focused primarily on one exposure at a time, which can underestimate the risk because the processes that cause extreme events are often interrelated and dependent on spatial and/or temporal factors. For example, the co-occurrence of heat waves and droughts is likely, as they are frequently connected at a physical level, with feedback loops between processes that can intensify both the drying and heating effects, leading to prolonged periods of hot and dry conditions. Health analyses focusing on a single climate-related exposure can substantially underestimate the risks associated with certain extremes if the impacts rely on multiple dependent climate threats. Adopting a multi-exposure approach, thus, is crucial to appropriately assess the impacts of climate extremes and to design effective adaptation and mitigation strategies.

Despite the demand from emerging climate change and health research, there is a lack of statistical and machine learning methodology tools to quantify the causal effects of multiple climate-relevant exposures. In most of the causal inference literature, the exposure of interest is often set as a univariate variable, either binary- or continuous-valued (**??**). The only exceptions are **?** and **?**, who proposed matching approaches in the context of multivariate continuous exposures: one relies on multivariate generalized propensity score (mvGPS) and the other implements matching on all

covariates directly. However, neither of them showed formal theoretical results for their proposed matching estimators.

In this article, we develop causal inference approaches to estimate the causal exposure-response relationship with theoretical guarantees. The main theoretical analyses focus on revisiting and extending the results of bias-corrected matching and double machine learning with univariate binary treatment for average treatment effect estimation (**??**). Specifically, we propose two bias-corrected matching estimators: the first one uses bias-corrected matching on all covariates, while the second one uses bias-corrected matching on estimated mvGPS, and both estimators have a random-forest mutation. Two estimators are based on debiased machine learning (**?**). In the first one, we regress Y on (W, X), while in the second one, we regress Y on W and the estimated mvGPS.

We use bias-corrected matching because it is a nonparametric method with root-N consistency, and it can avoid the problem of extreme value of the estimated mvGPS in weighting methods. Debiased machine learning is also a nonparametric method with root-n consistency, and we can improve the stability of the estimator against the extreme value of the estimated mvGPS by using the density ratio estimator in **?**. The idea of propensity score matching is used because propensity score matching leads to a variance reduction in the estimator, which is not shared by any other method, as in **?**. In our data simulations, we make comparisons among the four estimators and the TMLE estimator in **?**.

The proposed method is significant for the field of climate change and health. Although continuous causal inference is increasingly common in climate change research, our methods are among the earliest literature to discuss multivariate, potentially high-dimensional, exposure in a nonparametric or semiparametric model. Indeed, we are always exposed to a mixture of exposures that jointly influence our health in a non-additive way. For example, heat waves, wildfires, and droughts may well happen simultaneously, and synergize or antagonize the effect of one another in a complex physical process. Our methods can help researchers better deal with the non-additive non-linear effect of multivariate continuous exposures.

This article is organized as follows. In Section 2 we formally write the basic assumptions of our method. In Section 3 we describe the procedure and the theoretical properties of our estimators, which is the main analysis part of the article. In Section 4 we do data simulations under the settings of **?** to see the practical performance of our methods and traditional TMLE methods. In Section 5 we apply our methods to the data of Leng and output the dose-response curve of the medicine. In Section 6 we conclude our findings and make some discussions.

2 Problem Setup

In this article, we set the exposure variable W as a multivariate continuous random vector of the levels of the mediator variables between policy treatment and outcomes of interest. For example, when studying climate policy, the indices of climate change risks, such as floods and wildfire risks, can be set as the exposure variable (**?**). We also set a continuous or binary outcome Y and a set of covariates X are measured for N randomly samples. We are interested in the dose-response curve of the treatment intensity on a continuous scale. From the curve, we can see how the effect of the

treatment varies with its expected influence on the mediating variables.

Following **?**, we set the non-parametric structural equation model as follows:

$$X = f_X(U_X);\ W = f_W(X, U_W);\ Y = f_Y(W, X, U_Y) \tag{2.1}$$

where U_X, U_W and U_Y are exogenous random variables such that U_W is independent of U_Y, and U_X is independent of U_Y or U_W. In our setting, the treatment only influences U_W, keeping U_X and U_Y constant. We define the potential outcome as

$$Y(\mathbf{w}) = f_Y(\mathbf{w}, X, U_Y). \tag{2.2}$$

Suppose the policy turns the exposure variable of each treated unit from W to $W^{\mathbf{d}} = T(W; \mathbf{d})$, where T is an operator and $\mathbf{d}$ is a vector of the intensity of the policy. For example, an air-quality improvement policy may have the following influence on the exposure variable:

$$T_1(W; \mathbf{d}) = W - \mathbf{d}; \tag{2.3}$$

$$T_2(W; \mathbf{d}) = (W_1(1 - d_1), \cdots, W_{k_w}(1 - d_{k_w})). \tag{2.4}$$

Suppose the treatment is conducted on a unit i that $(W_i, X_i) \in \mathcal{D}_0$, where $\mathcal{D}_0$ is a known region. We then consider $E[Y(T(W; \mathbf{d})) - Y \mid (W, X) \in \mathcal{D}_0]$ as the causal effect of the treatment with intensity $\mathbf{d}$. When we construct the dose-response curve, we keep the function T constant and estimate the causal effect corresponding to treatment with different parameter d.

Extended from **?**, the main estimand in our analysis is

$$\theta(\mathbf{d}) = \mathrm{E}[Y(T(W; \mathbf{d})) - Y \mid (W, X) \in \mathcal{D}_0] \tag{2.5}$$

For simplicity, we'll substitute θ for $\theta(\mathbf{d})$ and substitute $T(W)$ for $T(W; \mathbf{d})$ below when we fix the value of $\mathbf{d}$ and calculate $\theta(\mathbf{d})$.

In some cases, we may not be aware of $\mathcal{D}$ and it is hard to set a reasonable $\mathcal{D}_0$, and there is an empirical way to set $\mathcal{D}_0$. Inspired by **?**, we can use the probability density ratio estimator in **?** to estimate $\frac{p(W=\mathbf{w}, X=\mathbf{x})}{p(W=T(\mathbf{w};\mathbf{d}), X=\mathbf{x})}$ and $\frac{p(W=T(\mathbf{w};\mathbf{d}), X=\mathbf{x})}{p(W=\mathbf{w}, X=\mathbf{x})}$. We suggest setting $\mathcal{D}_0$ as a subset of the region where the two probability density ratio estimators are both less than 50.

3 Methods

Assumption 3.1. *(i) Without policy intervention, $X \in R^{k_x}$, $W \in R^{k_w}$ and $Y \in R^{k_y}$ are continuously distributed. (ii) $\{Y(\mathbf{w})\}_{\mathbf{w}} \perp W \mid X$.*

Remark 3.1. In Pearl's model, we can view this assumption as: (i) The structural equation model is $X = f_X(U_X)$, $W = f_W(X, U_W)$ and $Y = f_Y(W, X, U_Y)$, where U_X, U_W and U_Y are exogenous random variables. U_W is independent of U_Y, and U_X is independent of U_Y or U_W. (ii) The policy of interest only affect U_W, keeping U_X and U_Y constant.

Assumption 3.2. *(i) Let the support of (W,X) be $\mathcal{D}$. $\mathcal{D}$ is a Cartesian product of compact intervals; (ii) The population treated by the policy are the units whose (W_i, X_i) are in a known region $\mathcal{D}_0 \subset \mathcal{D}$; (iii) $(W, X) \in \mathcal{D}_0 \Rightarrow (T(W; d), X) \in \mathcal{D}$; (iv) The transformation from (W, X) to $(T(W; \mathbf{d}), X)$ is a bijection from $\mathcal{D}_0$ to another region $\mathcal{D}_1$.*

Remark 3.2. Assumption 3.2 (i) is made to ensure the properties of series regression. This assumption is quite technical, so our results may still be valid even if it's violated.

We define the multivariate generalized propensity score (mvGPS) as follows:

$$e(\mathbf{w}, \mathbf{x}) = p(W = \mathbf{w}|X = \mathbf{x}) \tag{3.1}$$

As in regular causal inference, we make the overlap assumption:

Assumption 3.3. *$e(\mathbf{w}, \mathbf{x}) > \eta$ holds for any $(\mathbf{w}, \mathbf{x}) \in \mathcal{D}$, where $\eta > 0$ is an absolute constant.*

Remark 3.3. Under Assumption 3.3, $\mathcal{D}$ is of bounded measure. Therefore, it's important to make Assumption 3.2 (ii) in order to avoid the condition that $(T(W; d), X) \notin \mathcal{D}$.

Assumption 3.4. *Let $\mu(\mathbf{w}, \mathbf{x}) = \mathbb{E}[Y|W = \mathbf{w}, X = \mathbf{x}]$ and $\sigma^2(\mathbf{w}, \mathbf{x}) = \mathbb{E}\left[(Y - \mu(\mathbf{w}, \mathbf{x}))^2 \mid W = \mathbf{w}, X = \mathbf{x}\right]$. Then, (i) $\mu(\mathbf{w}, \mathbf{x})$ and $\sigma^2(\mathbf{w}, \mathbf{x})$ are Lipschitz continuous in $\mathcal{D}$, (ii) $\mathbb{E}[(Y(\mathbf{w}))^4|X = \mathbf{x}] \leq C$ for some finite C, for almost all $\mathbf{x}$, and (iii) $\sigma^2(\mathbf{w}, \mathbf{x})$ is bounded away from zero.*

4 Theory

In this section, we fix $\mathbf{d}$ and gives our estimators to estimate $\theta(\mathbf{d})$. For simplicity, we substitute θ for $\theta(\mathbf{d})$ and substitute $T(W)$ for $T(W; \mathbf{d})$. The following lemma is the foundation of our identification of θ:

Lemma 4.1. *Suppose $(\widetilde{W}, \widetilde{X})$ is of the same distribution of (W, X). Under Assumption 3.1 (i), (ii) and Assumption 3.2 (iv), we have*

$$\theta = \mathrm{E}\left[\mathrm{E}\left[Y|W = T(\widetilde{W}), X = \widetilde{X}\right]|(\widetilde{W}, \widetilde{X}) \in \mathcal{D}_0\right] - \mathrm{E}\left[Y|(W, X) \in \mathcal{D}_0\right] \tag{4.1}$$

$$\theta = \mathrm{E}\left[Y\frac{e(T^{-1}(W), X)}{e(W, X)}|(W, X) \in \mathcal{D}_1\right] - \mathrm{E}\left[Y|(W, X) \in \mathcal{D}_0\right] \tag{4.2}$$

$$\theta = \mathrm{E}\left[\mathrm{E}\left[Y|W = T(\widetilde{W}), \frac{e(T^{-1}(W), X)}{e(W, X)} = \frac{e(\widetilde{W}, \widetilde{X})}{e(T(\widetilde{W}), \widetilde{X})}\right] \mid (\widetilde{W}, \widetilde{X}) \in \mathcal{D}_0\right] - \mathrm{E}\left[Y|(W, X) \in \mathcal{D}_0\right] \tag{4.3}$$

Remark 4.1. The first formula is the intuition of the matching estimator on all covariates; The second formula is the intuition of the weighting estimator; The third formula is the intuition of the matching estimator on mvGPS. The three types of estimators, combined with the imputation estimator of θ, lead to the main results in our article.

According to the conditional term in the third formula above, we define a function $r(*) : \mathcal{D}_1 \to \mathbb{R}$ as

$$r(\mathbf{w}, \mathbf{x}) = \frac{e(T^{-1}(\mathbf{w}), \mathbf{x})}{e(\mathbf{w}, \mathbf{x})}. \tag{4.4}$$

Similar to **?**, we can derive the efficient influence curve of θ as in the following lemma.

Lemma 4.2. *Let $\xi_i(w,x) = \mathbb{1}((w,x) \in \mathcal{D}_i)$, the efficient curve of θ is*

$$D(P)(O) = \frac{(\mu(T(W),X) - \mu(W,X) - \theta)\xi_0(W,X) + (Y - \mu(W,X))(r(W,X)\xi_1(W,X) - \xi_0(W,X))}{P((W,X) \in \mathcal{D}_0)}, \tag{4.5}$$

and the efficiency bound in the estimation of θ is

$$\begin{aligned}
&\sigma^2 = E\left[D(P)(O)^2\right] \\
=&\frac{\mathrm{E}\left[((\mu(T(W),X) - \mu(W,X)) - \theta)^2 \mid (W,X) \in \mathcal{D}_0\right]}{P((W,X) \in \mathcal{D}_0)} \\
&+ \frac{\mathrm{E}\left[\sigma^2(W,X)(r(W,X)\xi_1(W,X) - \xi_0(W,X))^2\right]}{P((W,X) \in \mathcal{D}_0)^2} \\
=&\frac{\mathrm{E}\left[((\mu(T(W),X) - \mu(W,X)) - \theta)^2 \mid (W,X) \in \mathcal{D}_0\right]}{P((W,X) \in \mathcal{D}_0)} \\
&+ \frac{\mathrm{E}\left[\sigma^2(W,X)r^2(W,X) \mid (W,X) \in \mathcal{D}_1\right]\mathrm{P}((W,X) \in \mathcal{D}_1)}{P((W,X) \in \mathcal{D}_0)^2} \\
&+ \frac{\mathrm{E}\left[\sigma^2(W,X) \mid (W,X) \in \mathcal{D}_0\right]}{P((W,X) \in \mathcal{D}_0)} \\
&- \frac{2\mathrm{E}\left[\sigma^2(W,X)r(W,X) \mid (W,X) \in \mathcal{D}_0 \cap \mathcal{D}_1\right]P((W,X) \in \mathcal{D}_0 \cap \mathcal{D}_1)}{P((W,X) \in \mathcal{D}_0)^2}
\end{aligned}$$

4.1 Bias-corrected matching

Let $\mathcal{J}_M(i)$ represent the index set of the M nearest matches (W_j, X_j) of the unit $(T(W_i), X_i)$, among the units whose $(W,X) \in \mathcal{D}_1$. In other words, define

$$\mathcal{J}_M(i) := \Big\{ j : (W_j,X_j) \in \mathcal{D}_1, \sum_{k:(W_k,X_k)\in\mathcal{D}_1} \mathbb{1}\Big(\|(T(W_i),X_i) - (W_k,X_k)\| \le \|(T(W_i),X_i) - (W_j,X_j)\| \Big) \le M \Big\}.$$

Furthermore, introduce

$$K_M(\cdot) : \Big\{ i | (W_i,X_i) \in \mathcal{D}_1 \Big\} \to \mathbb{N}$$

to be the number of matched times of unit i, i.e.,

$$K_M(i) := \sum_{j:(W_j,X_j)\in\mathcal{D}_0} \mathbb{1}\left(\sum_{k:(W_k,X_k)\in\mathcal{D}_1} \mathbb{1}\Big(\|(T(W_j),X_j) - (W_k,X_k)\| \le \|(T(W_j),X_j) - (W_i,X_i)\| \Big) \le M \right).$$

Then, we can define the bias-corrected estimator as follows:

$$\widehat{\theta}_{bc} = \frac{1}{N_0} \sum_{p=1}^{2} \sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0} \frac{1}{M} \sum_{j\in\mathcal{J}_M(i)} (\widehat{\mu}(T(W_i),X_i) + Y_j - \widehat{\mu}(W_j,X_j) - Y_i)$$

$$= \frac{1}{N_0}\sum_{p=1}^{2}\left(\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}(\widehat{\mu}(T(W_i),X_i)-Y_i)+\frac{1}{N_0}\sum_{j\in\mathcal{D}_1\cap\mathcal{E}_1}\frac{K_M(j)}{M}(Y_j-\widehat{\mu}(W_j,X_j))\right)$$

where $\widehat{\mu}$ is the series estimator of μ. As in **?**, we make the following assumption:

Assumption 4.1. *(i)*

For the bias-corrected estimator, we have the following theorem:

Theorem 4.1. *Assume $M = O(N^v)$ for some $v < 1/2$ and $M \to \infty$ as $N \to \infty$. Under Assumption 3.1 to 3.4 and Assumption 4.1, we have*

$$\sqrt{N}(\widehat{\theta}_{bc}-\theta) \xrightarrow{d} N(0,\sigma^2) \tag{4.6}$$

Also, there's a consistent variance estimator

$$\widehat{\sigma}^2 = \frac{N}{N_0^2}\sum_{i\in\mathcal{D}_0}(\widehat{\mu}(T(W_i),X_i)-\widehat{\mu}(W_i,X_i))-\widehat{\theta})^2 \tag{4.7}$$

$$+\frac{N}{N_0^2}\left(\sum_{i\in\mathcal{D}_0}\widehat{\sigma}^2(W_i,X_i)+\sum_{i\in\mathcal{D}_1}\widehat{\sigma}^2(W_i,X_i)\frac{K_M^2(i)}{M^2}-2\sum_{i\in\mathcal{D}_0\cap\mathcal{D}_1}\widehat{\sigma}^2(W_i,X_i)\frac{K_M(i)}{M}\right) \tag{4.8}$$

Remark 4.2. Completely nonparametric. This estimator actually achieves the efficiency bound.

4.2 Propensity score matching

In this section, following **?**, we consider a generalized linear specification for the mvGPS, $e(\mathbf{w},\mathbf{x}) = F\left((\mathbf{w}^T,\mathbf{x}^T)\gamma^*\right)$, where F is a known function. Let $e(\mathbf{w},\mathbf{x};\gamma) = F\left((\mathbf{w}^T,\mathbf{x}^T)\gamma\right)$ and $r(\mathbf{w},\mathbf{x};\gamma) = \frac{e(\mathbf{w},\mathbf{x};\gamma)}{e(T^{-1}(\mathbf{w}),\mathbf{x};\gamma)}$. Let $\widehat{\gamma}_N$ be the maximum likelihood estimation (MLE) of γ^*. Consider a grid of cubes in $\mathbb{R}^{k_x+k_w}$ with sides of length $d/\sqrt{N}$, for arbitrary positive d . Let $\bar{\gamma}_N$ be the discretized estimator, defined as the midpoint of the cube $\widehat{\gamma}_N$ belongs to. If $\widehat{\gamma}_{N,j}$ is the jth component of the k-vector $\widehat{\theta}_N$, then the jth component of the k-vector $\bar{\theta}_N$ is $\bar{\theta}_{N,j} = (d/\sqrt{N})\left[\sqrt{N}\widehat{\theta}_{N,j}/d\right]$, where [.] is the nearest integer function. The bias-corrected mvGPS matching estimator is defined as follows:

$$\widehat{\theta}_{bc,mvGPS} = \frac{1}{N_0}\sum_{i\in\mathcal{D}_0}\frac{1}{M}\sum_{j\in\mathcal{J}_{M,\bar{\gamma}_N}(i)}(Y_j+\widehat{\bar{\mu}}(T(W_i),r(T(W_i),X_i;\bar{\gamma}_N))-\widehat{\bar{\mu}}(W_j,r(W_j,X_j;\bar{\gamma}_N))) \tag{4.9}$$

Here, $\mathcal{J}_{M,\gamma}(i)$ is the set of M nearest neighbors of unit i in $\mathcal{D}_1$ with regard to the distances between $r(W_i,X_i;\gamma)$ and $r(W_j,X_j;\gamma)$. $\widehat{\bar{\mu}}$ is the series estimator of $\bar{\mu}$. ($\bar{\mu}$ also changes with γ.) For this estimator, we have the following theorem

Theorem 4.2. *Assume $M = O(N^v)$ for some $v < 1/2$ and $M \to \infty$ as $N \to \infty$. Assume Assumption 3.1 to 3.4. Assume further that the support of $(W, r(W,X;\gamma))$ is the Cartesian product of compact intervals and that Assumption 4 and 5 in **?** hold. Then, we have*

$$\sqrt{N_0}(\widehat{\theta}_{bc,mvGPS}-\theta) \xrightarrow{d} N(0,\sigma^2_{GPS}-\mathrm{P}_0 c^T I_F^{-1} c) \tag{4.10}$$

where

$$\sigma^2_{GPS} = \mathrm{E}\left[(\bar{\mu}(T(W), r(T(W), X)) - \theta)^2 \mid (W, X) \in \mathcal{D}_0\right] + \mathrm{E}\left[\bar{\sigma}^2(T(W), r(T(W), X))r(T(W), X) \mid (W, X) \in \mathcal{D}_0\right] \tag{4.11}$$

$$\mathrm{P}_0 = \mathrm{P}((W, X) \in \mathcal{D}_0) \tag{4.12}$$

$$c = \mathrm{E}\left[\underset{(W,X)\in\mathcal{D}_0}{\mathrm{Cov}}(\mu(W, X), X|(W^T, X^T)\gamma^*)\frac{f(W, X; \gamma^*)}{e(T^{-1}(W), X; \gamma^*)} \mid (W, X) \in \mathcal{D}_1\right] \tag{4.13}$$

$$I_F = \mathrm{E}\left[\frac{f(W, X; \gamma^*)^2}{e(W, X; \gamma^*)^2}\begin{pmatrix} W \\ X \end{pmatrix}(W^T, X^T)\right] \tag{4.14}$$

Remark 4.3. Semiparametric. If $W|X$ is Gaussian linear, the parametric specification is correct.

4.3 Double machine learning

Let $O = (Y, W, X)$. We exploit the following algorithm to get the $\sqrt{N}$-consistent estimation of θ.

Definition 4.1. DML(Ψ): Input a Neyman-orthogonal score $\Psi(O; \theta, \eta)$, where $\eta = (r, \mu)$, the nuisance parameter. Then (1), For the sample $\{O_i\}_{i=1}^N$, randomly partition the sample whose $(W_i, X_i) \in \mathcal{D}_0$ into folds $(I_\ell)_{\ell=1}^L$ of approximately equal size. Denote by I_ℓ^c the complement of I_ℓ in $\{O_i\}_{i=1}^N$. (2) For each ℓ , estimate $\widehat{\eta}_\ell = (\widehat{r}_\ell, \widehat{\mu}_\ell)$ from observations in I_ℓ^c . (3) Estimate θ as a root of: $0 = \frac{1}{N}\sum_{\ell=1}^L \sum_{i\in I_\ell} \Psi\left(O_i; \theta, \widehat{\eta}_\ell\right)$. Output $\widehat{\theta}$ and the estimated scores $\widehat{\Psi}^o\left(O_i\right) = \Psi\left(O_i; \widehat{\theta}, \widehat{\eta}_\ell\right)$ for each $i \in I_\ell$ and each ℓ .

Therefore the estimator is defined as

$$\widehat{\theta}_s := \mathrm{DML}\left(\Psi\right) \tag{4.15}$$

for the score

$$\Psi(O; \theta, \eta) = (Y - \mu(W, X))r(W, X) + \mu(T(W), X) - \theta \tag{4.16}$$

$$or\ \Psi(O; \theta, \eta) = (Y - \bar{\mu}(W, r(W, X)))r(W, X) + \bar{\mu}(T(W), r(T(W), X)) - \theta \tag{4.17}$$

which is orthogonal at the true values of nuisance parameters. We can use regular regression methods to get the initial estimation of μ and use the method in **?** to get the initial estimation of r. We have the following theorem:

Theorem 4.3. *Suppose that Ψ and the machine learners $\widehat{\eta}_\ell$ of η_0 obey Assumptions 3.1 and 3.2 in* **?***, in particular the estimators $\widehat{\eta}_l \in \mathcal{H}_s^2 \times \mathcal{Q}_s^2$ has the convergence rate of*

$$\|\widehat{mu}_\ell - mu\|\|\widehat{r}_\ell - r\| = o_P(n^{-1/2}) \tag{4.18}$$

$$\tag{4.19}$$

Then the estimator is asymptotically linear and Gaussian with the influence function:

$$\Psi^o(O) := \Psi_\theta\left(O; \theta, \eta_0\right) \tag{4.20}$$

The covariance of the scores can be estimated by the empirical analogues using the covariance of the estimated scores.

Remark 4.4. Nonparametric. The asymptotic variances under the two scores are same to σ^2 and σ^2_{GPS} in the matching part respectively.

5 Simulation

6 Application

7 Conclusion

8 Proofs of the main results

Proof of Lemma 4.1. Since

$$\mathrm{E}[Y(T(W)) \mid W = w, X = x] = \mathrm{E}[Y(T(w)) \mid W = T(w), X = x] = \mathrm{E}[Y \mid W = T(w), X = x],$$

we know that

$$\mathrm{E}[Y(T(W)) \mid (W, X) \in \mathcal{D}_0] = \mathrm{E}\left[\mathrm{E}\left[Y|W = T(\widetilde{W}), X = \widetilde{X}\right] |(\widetilde{W}, \widetilde{X}) \in \mathcal{D}_0\right]$$

and that

$$\theta = \mathrm{E}\left[\mathrm{E}\left[Y|W = T(\widetilde{W}), X = \widetilde{X}\right] |(\widetilde{W}, \widetilde{X}) \in \mathcal{D}_0\right] - \mathrm{E}[Y \mid (W, X) \in \mathcal{D}_0].$$

Based on (4.1), we can derive (4.2) as follows:

$$\begin{aligned}
\theta =& \int_{(w,x)\in\mathcal{D}_0} \mathrm{E}\left[Y|W = T(w), X = x\right] \frac{p_{W,X}(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} dwdx - \mathrm{E}[Y \mid (W, X) \in \mathcal{D}_0] \\
=& \int_{(w,x)\in\mathcal{D}_1} \mathrm{E}\left[Y|W = w, X = x\right] \frac{p_{W,X}(T^{-1}(w), x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} dwdx - \mathrm{E}[Y \mid (W, X) \in \mathcal{D}_0] \\
=& \int_{(w,x)\in\mathcal{D}_1} \mathrm{E}\left[Y|W = w, X = x\right] \frac{r(w, x)p_{W,X}(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} dwdx - \mathrm{E}[Y \mid (W, X) \in \mathcal{D}_0] \\
=& \mathrm{E}\left[Y r(W, X) \frac{\mathrm{P}((W, X) \in \mathcal{D}_1)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} |(W, X) \in \mathcal{D}_1\right] - \mathrm{E}[Y \mid (W, X) \in \mathcal{D}_0].
\end{aligned}$$

To prove (4.3), we first notice that

$$\frac{p_{W|(X,r(T(w),X))}(w \mid x, r(T(w), x))}{p_{W|(X,r(T(w),X))}(T(w) \mid x, r(T(w), x)} = \frac{p_{W|r(T(w),X)}(w \mid r(T(w), x))}{p_{W|r(T(w),X)}(T(w) \mid r(T(w), x))} = r(T(w), x).$$

Given this result, we have

$$\begin{aligned}
&\mathrm{E}[Y(T(W)) \mid W = w, r(T(W), X) = r(T(w), x)] \\
=&\mathrm{E}[Y(T(w)) \mid W = w, r(T(w), X) = r(T(w), x)] \\
=& \int_{r(T(w),X)=r(T(w),x)} E[Y(T(w)) \mid W = w, X = x] p_{X|(W,r(T(w),X))}(x \mid w, r(T(w), x))dx \\
=& \int_{r(T(w),X)=r(T(w),x)} E[Y \mid W = T(w), X = x] p_{X|(W,r(T(w),X))}(x \mid w, r(T(w), x))dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_{r(T(w),X)=r(T(w),x)} E[Y \mid W = T(w), X = x] \frac{p_{(X,W)|r(T(w),X)}(x, w \mid r(T(w), x))}{p_{W|r(T(w),X)}(w \mid r(T(w), x))} dx \\
&= \int_{r(T(w),X)=r(T(w),x)} E[Y \mid W = T(w), X = x] \frac{p_{X|r(T(w),X)}(x \mid r(T(w), x)) p_{W|(X,r(T(w),X))}(w \mid x, r(T(w), x))}{p_{W|r(T(w),X)}(w \mid r(T(w), x))} dx \\
&= \int_{r(T(w),X)=r(T(w),x)} E[Y \mid W = T(w), X = x] \frac{p_{X|r(T(w),X)}(x \mid r(T(w), x)) p_{W|(X,r(T(w),X))}(T(w) \mid x, r(T(w), x))}{p_{W|r(T(w),X)}(T(w) \mid r(T(w), x))} dx \\
&= \int_{r(T(w),X)=r(T(w),x)} E[Y \mid W = T(w), X = x] p_{X|(W,r(T(w),X))}(x \mid T(w), r(T(w), x)) dx \\
&= \int_{r(T(w),X)=r(T(w),x)} E[Y \mid W = T(w), X = x] p_{X|(W,r(W,X))}(x \mid T(w), r(T(w), x)) dx \\
&= \mathrm{E}\left[Y | W = T(w), r(W, X) = r(T(w), x)\right].
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\theta =& \mathrm{E}\left[\mathrm{E}\left[Y(T(W)) | W = \widetilde{W}, r(T(W), X) = r(T(\widetilde{W}), \widetilde{X})\right] | (\widetilde{W}, \widetilde{X}) \in \mathcal{D}_0\right] - \mathrm{E}[Y \mid (W, X) \in \mathcal{D}_0] \\
=& \mathrm{E}\left[\mathrm{E}\left[Y | W = T(\widetilde{W}), r(W, X) = r(T(\widetilde{W}), \widetilde{X})\right] \mid (\widetilde{W}, \widetilde{X}) \in \mathcal{D}_0\right] - \mathrm{E}\left[Y | (W, X) \in \mathcal{D}_0\right].
\end{aligned}
$$

□

Proof of Lemma 4.2. In calculating the variance bounds of θ, we follow the method in Section 3.3 of **?**. First, the tangent space is characterized. Consider the marginal distribution of (W, X) and the conditional distribution of Y, we can find the tangent space $\mathscr{S}$ of (Y, W, X) to be

$$
\mathscr{S} = \left\{ s(y, w, x) = f_1(w, x) + f_2(y, w, x) \right\},
$$

where $\int f_2(y, w, x) p_{Y|W,X}(y \mid w, x) dy = 0$ *a.s.* for $(w, x) \in \mathcal{D}$ and that $\int f_1(w, x) p_{W,X}(w, x) dw dx = 0$. For any $s(y, w, x) \in \mathcal{S}$, we can decompose it to be $f_1(w, x)$ and $f_2(y, w, x)$ as above.

For any regular parametric submodel with parameter γ, we find from (4.2) that

$$
\begin{aligned}
\theta(\gamma) = & \int \left(\int y p_{Y|W,X}(y \mid w, x, \gamma) dy \right) \frac{r(w, x \mid \gamma) \xi_1(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} p_{W,X}(w, x \mid \gamma) dw dx \\
& - \iint \frac{y \xi_0(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} p_{Y,W,X}(y, w, x \mid \gamma) dy dw dx.
\end{aligned}
$$

Then, we can find that θ is pathwise differentiable to γ, and we can derive the following equation if we denote γ_0 as the parameter for the true model:

$$
\begin{aligned}
\frac{\partial \theta(\gamma_0)}{\partial \gamma} = & \int \left(\int y f_2(y, w, x) p_{Y|W,X}(y \mid w, x) dy \right) \frac{r(w, x) \xi_1(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} p_{W,X}(w, x) dw dx \\
& + \int \left(\int y p_{Y|W,X}(y \mid w, x) dy \right) \frac{r(w, x) \xi_1(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} f_1(w, x) p_{W,X}(w, x) dw dx \\
& + \int \left(\int y p_{Y|W,X}(y \mid w, x) dy \right) \frac{r(w, x)(f_1(T^{-1}(w), x) - f_1(w, x)) \xi_1(w, x)}{\mathrm{P}((W, X) \in \mathcal{D}_0)} p_{W,X}(w, x) dw dx
\end{aligned}
$$

$$
\begin{aligned}
&- \iint \frac{y\xi_0(w,x)}{\mathrm{P}((W,X)\in\mathcal{D}_0)} s(y,w,x) p_{Y,W,X}(y,w,x)dwdx \\
=&\mathrm{E}[(f_1(W,X)+f_2(Y,W,X))\frac{r(W,X)\xi_1(W,X)}{\mathrm{P}((W,X)\in\mathcal{D}_0)}(Y-\mu(W,X))] \\
&+\mathrm{E}[(f_1(W,X)+f_2(Y,W,X))\frac{r(W,X)\xi_1(W,X)}{\mathrm{P}((W,X)\in\mathcal{D}_0)}\mu(W,X)] \\
&-\mathrm{E}[(f_1(W,X)+f_2(Y,W,X))\frac{r(W,X)\xi_1(W,X)}{\mathrm{P}((W,X)\in\mathcal{D}_0)}\mu(W,X)] \\
&+\int \mu(w,x)\frac{f_1(T^{-1}(w),x)\xi_1(w,x)}{\mathrm{P}((W,X)\in\mathcal{D}_0)} p_{W,X}(T^{-1}(w),x)dwdx \\
&-\mathrm{E}[s(Y,W,X)\frac{\xi_0(W,X)}{\mathrm{P}((W,X)\in\mathcal{D}_0)}(Y-\theta)] \\
=&\mathrm{E}\left[s(Y,W,X)\frac{r(W,X)\xi_1(W,X)(Y-\mu(W,X))-\xi_0(W,X)(Y-\theta)}{\mathrm{P}((W,X)\in\mathcal{D}_0)}\right] \\
&+\int \mu(T(w),x)\frac{f_1(w,x)\xi_0(w,x)}{\mathrm{P}((W,X)\in\mathcal{D}_0)} p_{W,X}(w,x)dwdx \\
=&\mathrm{E}\left[s(Y,W,X)\frac{r(W,X)\xi_1(W,X)(Y-\mu(W,X))-\xi_0(W,X)(Y-\theta)-\xi_0(W,X)\mu(T(W),X)}{\mathrm{P}((W,X)\in\mathcal{D}_0)}\right] \\
=&\mathrm{E}\left[s(Y,W,X)D(P)(O)\right]
\end{aligned}
$$

Finally, since $D(P)(O)\in\mathscr{S}$ holds, $D(P)(O)$ is the efficient curve of θ and $\sigma^2=\mathrm{E}[D(P)(O)^2]$ is the efficiency bound for asymptotic variance. □

Proof of Theorem 4.1. For each sample splitting scheme, let $\mathcal{E}_0$ be the set where W is changed and let $\mathcal{E}_1$ be the set where W is fixed. Then, the estimator obtained with the two complementary sample splitting schemes can be decomposed as follows:

$$
\begin{aligned}
&N^{1/2}(\widehat{\theta}_{bc}-\theta) \\
=&\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\{[\widehat{\mu}(T(W_i),X_i)-\mu(T(W_i),X_i)]-[\widehat{\mu}(W_j,X_j)-\mu(W_j,X_j)]\} \\
&+\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}(\mu(T(W_i),X_i)-\mu(W_i,X_i)-\theta) \\
&+\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i=1}^{N}(\mathbb{1}(i\in\mathcal{D}_1\cap\mathcal{E}_1)r(W_i,X_i)-\mathbb{1}(i\in\mathcal{D}_0\cap\mathcal{E}_0))\epsilon_i \\
&+\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i=1}^{N}\mathbb{1}(i\in\mathcal{D}_1\cap\mathcal{E}_1)(\frac{K_M(i)}{M}-r(W_i,X_i))\epsilon_i.
\end{aligned}
$$

Among the four terms, notice that the middle two terms are asymptotically independent, we only

need to prove the following four statements:

$$\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\{[\widehat{\mu}(T(W_i),X_i)-\mu(T(W_i),X_i)]-[\widehat{\mu}(W_j,X_j)-\mu(W_j,X_j)]\}=o_P(1) \tag{8.1}$$

$$\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}(\mu(T(W_i),X_i)-\mu(W_i,X_i)-\theta)\xrightarrow{d}N(0,\sigma_1^2) \tag{8.2}$$

$$\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i=1}^{N}(\mathbb{1}(i\in\mathcal{D}_1\cap\mathcal{E}_1)r(W_i,X_i)-\mathbb{1}(i\in\mathcal{D}_0\cap\mathcal{E}_0))\epsilon_i\xrightarrow{d}N(0,\sigma_2^2). \tag{8.3}$$

$$\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i=1}^{N}\mathbb{1}(i\in\mathcal{D}_1\cap\mathcal{E}_1)(\frac{K_M(i)}{M}-r(W_i,X_i))\epsilon_i=o_P(1) \tag{8.4}$$

To prove (8.1), we follow the proof of Lemma C.3 in **?**. Suppose $j_m(i)$ is the $m-th$ nearest neighbor of i. Suppose $U_{m,i}=(W_{j_m(i)},X_{j_m(i)})-(T(W_i),X_i)$. Suppose ∂^t and $U_{m,i}^t$ can be defined with vector t, where t indicates the order of each variable/component. Suppose Λ_ℓ is the set of all nonnegative integer vectors whose sum is ℓ. Let $k_w+k_x=d$ and $k=\lfloor d/2\rfloor+1$. From Tyler expansion, we know that

$$\left|\mu(W_{j_m(i)},X_{j_m(i)})-\mu(T(W_i),X_i)-\sum_{\ell=1}^{k-1}\frac{1}{\ell!}\sum_{t\in\Lambda_\ell}\partial^t\mu(T(W_i),X_i)U_{m,i}^t\right|\le\max_{t\in\Lambda_k}\left\|\partial^t\mu\right\|_\infty\frac{1}{k!}\sum_{t\in\Lambda_k}\|U_{m,i}\|^k,$$

$$\left|\widehat{\mu}(W_{j_m(i)},X_{j_m(i)})-\widehat{\mu}(T(W_i),X_i)-\sum_{\ell=1}^{k-1}\frac{1}{\ell!}\sum_{t\in\Lambda_\ell}\partial^t\widehat{\mu}(T(W_i),X_i)U_{m,i}^t\right|\le\max_{t\in\Lambda_k}\left\|\partial^t\widehat{\mu}\right\|_\infty\frac{1}{k!}\sum_{t\in\Lambda_k}\|U_{m,i}\|^k,$$

and

$$\sum_{\ell=1}^{k-1}\left|\frac{1}{\ell!}\sum_{t\in\Lambda_\ell}(\partial^t\widehat{\mu}(T(W_i),X_i)-\partial^t\mu(T(W_i),X_i))U_{m,i}^t\right|\le\sum_{\ell=1}^{k-1}\max_{t\in\Lambda_\ell}\left\|\partial^t\widehat{\mu}-\partial^t\mu\right\|_\infty\frac{1}{\ell!}\sum_{t\in\Lambda_\ell}\|U_{m,i}\|^\ell.$$

So,

$$\begin{aligned}&\left|\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\{[\widehat{\mu}(T(W_i),X_i)-\mu(T(W_i),X_i)]-[\widehat{\mu}(W_j,X_j)-\mu(W_j,X_j)]\}\right|\\ \le&N^{1/2}\left(\max_{\omega\in\{0,1\},t\in\Lambda_k}\left\|\partial^t\mu\right\|_\infty+\max_{\omega\in\{0,1\},t\in\Lambda_k}\left\|\partial^t\widehat{\mu}\right\|_\infty\right)\left(\frac{1}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}\|U_{m,i}\|^k\right)\\ &+N^{1/2}\sum_{\ell=1}^{k-1}\left(\max_{\omega\in\{0,1\},t\in\Lambda_\ell}\left\|\partial^t\widehat{\mu}-\partial^t\mu\right\|_\infty\right)\left(\frac{1}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}\|U_{m,i}\|^\ell\right)\end{aligned}$$

With the Lemma C.2 in **?**, we know that for any integer $p>0$,

$$\frac{1}{n}\sum_{i=1}^{n}\|U_{m,i}\|^p=O_p\left(\left(\frac{M}{|\mathcal{D}_1\cap\mathcal{E}_1|}\right)^{p/d}\right)=O_p\left(\left(\frac{M}{N_1}\right)^{p/d}\right).$$

Therefore, with the assumptions on the convergence rate of $\widehat{\mu}$ and the derivative of the μ function, we know that

$$\begin{aligned}
&\left|\frac{N^{1/2}}{N_0}\sum_{p=1}^{2}\sum_{i\in\mathcal{D}_0\cap\mathcal{E}_0}\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)}\{[\widehat{\mu}(T(W_i),X_i)-\mu(T(W_i),X_i)]-[\widehat{\mu}(W_j,X_j)-\mu(W_j,X_j)]\}\right| \\
=&N^{1/2}\left(O_p(1)O_p\left(\left(\frac{M}{N_1}\right)^{k/d}\right)+\max_{\ell\in[k-1]}O_p(n^{-\gamma_\ell})O_p\left(\left(\frac{M}{N_1}\right)^{\ell/d}\right)\right) \\
=&N^{1/2}O_p\left(\left(\frac{M}{N_1}\right)^{k/d}+\max_{\ell\in[k-1]}n^{-\gamma_\ell}\left(\frac{M}{N_1}\right)^{\ell/d}\right)=o_P(1).
\end{aligned}$$

As for (8.2) and (8.3), it's easy to verify that

$$(8.2)=\frac{N^{1/2}}{N_0}\sum_{i=1}^{N}\mathbb{1}(i\in\mathcal{D}_0)(\mu(T(W_i),X_i)-\mu(W_i,X_i)-\theta)\xrightarrow{d}N(0,\sigma_1^2)$$

and

$$(8.3)=\frac{N^{1/2}}{N_0}\sum_{i=1}^{N}(\mathbb{1}(i\in\mathcal{D}_1)r(W_i,X_i)-\mathbb{1}(i\in\mathcal{D}_0\cap\mathcal{E}_0))\epsilon_i\xrightarrow{d}N(0,\sigma_2^2).$$

As for (8.4), similar to the proof of Lemma C.1 in **?**, we can apply the Linderberg-Feller central limit theorem to prove that

$$\frac{N^{1/2}}{VN_0}\sum_{i=1}^{N}\sum_{p=1}^{2}\mathbb{1}(i\in\mathcal{D}_1\cap\mathcal{E}_1)(\frac{K_M(i)}{M}-r(W_i,X_i))\epsilon_i\xrightarrow{d}N(0,1),$$

where

$$\begin{aligned}
V=&\frac{N}{N_0^2}\sum_{i=1}^{N}Var\left(\sum_{p=1}^{2}\mathbb{1}(i\in\mathcal{D}_1\cap\mathcal{E}_1)(\frac{K_M(i)}{M}-r(W_i,X_i))\epsilon_i\mid W_i,X_i\right) \\
=&\frac{N}{N_0^2}\sum_{i=1}^{N}\mathbb{1}(i\in\mathcal{D}_1)(\frac{K_M(i)}{M}-r(W_i,X_i))^2\sigma^2(W_i,X_i).
\end{aligned}$$

From Theorem B.2 in **?**, we know that

$$EV\leq\frac{NN_1C}{N_0^2}E\left(\frac{K_M(i)}{M}-r(W_i,X_i)\right)^2=o\left(\frac{NN_1C}{N_0^2}\right)=o(1),$$

where C is the upper bound of $\sigma^2(W_i,X_i)$. Therefore, (8.4)$=O_P(EV)=o_P(1)$. □

Proof of Theorem 4.2. First, we can consider the case when the estimator $\widehat{\gamma}$ happens to be the true γ. In this case, the asmptotic normality of our matching estimator follows from the proof of Theorem 4.1, and we can derive the influence function of the estimator. Second, we can calculate the covariance between the estimator and the score function of γ. Finally, we can apply Le Cam's third lemma on the discretized values of $\widehat{\gamma}$ to finish the proof. □