RESEARCH ARTICLE

Causal effect estimation for multivariate continuous treatments

Juan Chen  |  Yingchun Zhou

Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, School of Statistics, Shanghai, People's Republic of China

Correspondence
Yingchun Zhou, Key Laboratory of Advanced Theory and Application in Statistics and Data Science-MOE, School of Statistics, East China Normal University, Shanghai, 200062, People's Republic of China.
Email: yczhou@stat.ecnu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Numbers: 11771146, 11831008; Program of Shanghai Subject Chief Scientist, Grant/Award Number: 14XD1401600; 111 Project, Grant/Award Number: B14019

This article has earned an open data badge "Reproducible Research" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract
Causal inference is widely used in various fields, such as biology, psychology, and economics, etc. In observational studies, balancing the covariates is an important step in estimating the causal effect. This study extends the one-dimensional entropy balancing method to multiple dimensions to balance the covariates. Both parametric and nonparametric methods are proposed to estimate the causal effect of multivariate continuous treatments and theoretical properties of the two estimations are provided. Furthermore, the simulation results show that the proposed method is better than other methods in various cases. Finally, the proposed method is applied to analyze the impact of the duration and frequency of smoking on medical expenditure. The results from the parametric method indicate that the frequency of smoking increases medical expenditure while the duration of smoking does not. The results from the nonparametric method indicate that there is a short-term downward trend and then a long-term upward trend as the duration and frequency of smoking increase.

KEYWORDS
causal inference, causal effect, entropy balancing, multivariate continuous treatments

1  |  INTRODUCTION

For decades, causal inference has been widely used in many fields, such as biology, psychology, and economics, etc. Most of the current research is based on univariate treatment (binary treatment, multivalued treatment, continuous treatment) (Chan et al., 2016; Dong et al., 2021; Fong et al., 2018; Hsu et al., 2022; Imai & Ratkovic, 2014; Xiong et al., 2017; Yiu & Su, 2018; Zhu et al., 2015; Zubizarreta, 2015a). Some research is focused on multivariate categorical treatments, such as factorial designs and conjoint analysis that estimate the main or interaction effect of any combination level of treatments (Dasgupta et al., 2015; Hainmueller et al., 2014). However, sometimes decision-makers are interested in the causal effects of multivariate continuous treatments in real life. For example, when considering the impact of the export and import volume on a country’s GDP, one is interested in a bivariate continuous treatment. The methods for multivariate categorical treatments are not suitable for multivariate continuous treatments, and there have been few research on this topic. The goal of this paper is to develop a new method to estimate the causal effect function for multivariate continuous treatments.
A major challenge for inferring the causal effect in observational studies is to balance the confounding covariates, which affect both the treatment and outcome variables. The covariate balancing propensity score method is widely used in controlling for confounding (Hirano & Imbens, 2004; Rosenbaum & Rubin, 1983, 1984, 1985; Robins et al., 2000). When using the parametric method to model the propensity score, the estimation bias will be large if the model is misspecified. Therefore, some nonparametric methods for estimating the propensity score have been proposed, such as the kernel density estimation (Robbins et al., 2020). In addition, in recent years, some studies have used optimized weighting methods to directly optimize the balancing of covariates (Hainmueller, 2012; Imai & Van Dyk, 2004; Vegetabile et al., 2020). These methods avoid the direct construction of the propensity scores; therefore, the obtained estimates achieve higher robustness. One of the methods, the entropy-balancing method, has been established as being doubly robust, in that a consistent estimate can be obtained when one of the two models, either the treatment assignment model or the outcome model, is correctly specified (Zhao & Percival, 2017). Furthermore, this method can be easily implemented by solving a convex optimization problem. To this end, Ai et al. (2021) extended this method to univariate continuous treatment to achieve exact balancing of the covariates. However, the balancing conditions cannot hold exactly when the dimension of covariates or treatments is high. Particularly, for multivariate treatments, it is difficult for the balancing conditions to hold exactly even when the dimensions of treatments and covariates are low. To meet this challenge, the literature has shown that approximate balance can trade bias for variance and it works well in practice in both low- and high-dimensional settings (Athey et al., 2018; Wang & Zubizarreta, 2020). Therefore, we extend this method to multivariate continuous treatments to balance the covariates approximately in this study.

This study has the following three contributions: First, it extends the univariate entropy-balancing method to multivariate continuous treatments to balance the covariates approximately. Second, both parametric and nonparametric causal effect estimation methods for multivariate continuous treatments are proposed. Under the parametric framework, a weighted optimization estimation is defined and its theoretical properties are provided. Under the nonparametric framework, B-splines are used to approximate the causal effect function, and the convergence rate of the estimation is provided. Third, we apply the proposed method to explore the impact of the duration and frequency of smoking on medical expenditure. The results from the parametric method show that the frequency of smoking increases medical costs significantly. The results from the nonparametric method show that there is a short-term downward trend and then a long-term upward trend when both the duration and frequency of smoking increase.

The remainder of this paper is organized as follows: In Section 2, we introduce the motivating example in this study. In Section 3, the entropy balancing for multivariate treatment (EBMT) method is proposed. In Section 4, theoretical properties of the parametric and nonparametric estimation are shown. In Section 5, methods for variance estimation and confidence interval (CI) construction are provided. In Section 6, a numerical simulation is performed to evaluate the properties of the EBMT method. In Section 7, the EBMT method is applied to the real data analysis. The conclusions and discussions are summarized in Section 8.

# Motivating Example

In this section, we introduce an observational study that contains two-dimensional treatments that motivates our methodology. The causal relationship between smoking and medical expenditure has long been a hot topic of research. Most studies estimate the causal effect of a binary treatment (smoker/nonsmoker) on medical costs (Larsen, 1999; Rubin, 2000; Zeger et al., 2000). However, the binary treatment is too incapable of describing the status of a smoker. A better way is to regard the duration and frequency of smoking as bivariate continuous treatments and study their causal effect on medical costs, which motivates this study.

The data we used are extracted from the 1987 National Medical Expenditure Survey (NMES), which includes the detailed information about the duration and frequency of smoking and is originally studied by Johnson et al. (2003). Furthermore, Imai and Dyk (2004) directly estimate the causal effect of smoking on medical expenditures with a univariate continuous treatment packyear, which is defined as

\[
\text{packyear} = \frac{\text{Number of cigarettes per day}}{20} \times \text{Number of years smoked}.
\]

Since this method can only obtain the impact of cumulative smoking on medical expenditures, the authors also conduct an analysis with a bivariate continuous treatment (the duration and frequency of smoking). Specifically, it first estimates
the propensity functions of the duration and frequency of smoking, respectively. Then it constructs subclasses based on the two estimated propensity functions and estimates the causal effects within each subclass using a linear regression. Finally, the overall average effects of the duration and frequency are obtained by taking the weighted average of the within-subclass treatment effects. However, this method roughly subclassifies the data and estimates the propensity functions of the duration and frequency separately, which ignores the relationship between these two variables.

Motivated by this example, we develop a method to estimate the causal effect jointly when the treatment is multivariate.

3 ENTROPY BALANCING FOR MULTIVARIATE TREATMENTS

In this section, we introduce the EBMT method to obtain weights that can balance covariates and both parametric and nonparametric approaches are developed to estimate the causal effect function.

3.1 Notation and assumptions

Suppose that the treatment for subject $i$ is $T_i$, whose support is $T \subset R^p$. $X_i \in R^q$ denotes the observed covariates, where $p, q$ denote the dimensions of the treatments and covariates, respectively. Suppose that for each subject, there exists a potential outcome $Y_i(t)$ for all $t \in T$. The observed outcome is defined as $Y_i = Y_i(t)$ if $T_i = t$. Assume that a sample of observations $\{Y_i, T_i, X_i\}$ for $i \in \{1, \ldots, n\}$ is independently drawn from a joint distribution $f(Y, T, X)$. For notational convenience, the treatments and covariates are assumed to be standardized.

To perform causal inference with the observational data, the following three standard assumptions are made (Hirano & Imbens, 2004; Imai & Van Dyk, 2004).

Assumption 1 (Strong Ignorability). $T_i \perp Y_i(t) | X_i$, which means that the treatment assignment is independent of the counterfactual outcomes, conditional on covariates. This implies that there is no unmeasured confounding.

Assumption 2 (Positivity). $f_{T|X}(T_i = t | X_i) > 0$ for all $t \in T$, and the conditional density $f(T_i | X_i)$ is called the generalized propensity score (Imbens, 2000).

Assumption 3 (Stable Unit Treatment Value Assumption (SUTVA)). Assume that there is no interference among the units, which means that each individual’s outcome depends only on their own level of treatment intensity.

The goal of this paper is to estimate the causal effect function $E(Y(t))$ for a multivariate continuous treatment $T$ based on the above assumptions. In order to estimate the causal effect function with observational data, we first define the stabilized weight as

$$w_i = \frac{f(T_i)}{f(T_i | X_i)},$$

then under the strong ignorability assumption, one can estimate the causal effect function based on the stabilized weight by using a parametric or nonparametric method.

3.2 Entropy balancing for multivariate treatments

The entropy-balancing method (Hainmueller, 2012) is used to determine the optimal weight for inferring causal effects. It has been used for univariate treatment, and here we extend this method to multivariate treatments and to balance covariates approximately.

Note that the stabilized weight

$$w_i = \frac{f(T_i)}{f(T_i | X_i)},$$

(1)
satisfies the following conditions for any suitable functions $\alpha(T)$ and $\beta(X)$:

$$
E(w_i \alpha(T_i) \beta(X_i)') = \int \left\{ \int f(T_i | X_i) \alpha(T_i) \beta(X_i)' f(T_i, X_i) dT_i dX_i \right\} \beta(X_i)' f(X_i) dX_i
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \alpha_{K1,l}(T_i) \beta_{K2,\bar{l}}(X_i) - \left( \frac{1}{n} \sum_{i=1}^{n} \alpha_{K1,l}(T_i) \right) \left( \frac{1}{n} \sum_{i=1}^{n} \beta_{K2,\bar{l}}(X_i) \right) \leq \lambda_{l,\bar{l}},
$$

(2)

However, Equation (2) implies an infinite number of moment conditions, which is impossible to be solved with a finite sample of observations. Hence, the finite-dimensional sieve space is considered to approximate the infinite-dimensional function space. Specifically, let

$$
\alpha_{K1}(T) = (\alpha_{K1,1}(T), \alpha_{K1,2}(T), \ldots, \alpha_{K1,K1}(T))',
$$

$$
\beta_{K2}(X) = (\beta_{K2,1}(X), \beta_{K2,2}(X), \ldots, \beta_{K2,K2}(X))'
$$

denote the known basis functions, then

$$
E(w_i \alpha_{K1}(T_i) \beta_{K2}(X_i)') = E(\alpha_{K1}(T_i)) E(\beta_{K2}(X_i)').
$$

(3)

Actually, this framework can be generalized to multivariate treatments with categorical variables or combinations of categorical and continuous variables, which is beyond the scope of this article.

In practice, the covariate balancing conditions given in Equation (3) cannot hold exactly with high-dimensional covariates or treatments. Particularly, for multivariate treatments, it is difficult to hold exactly even when the dimensions of treatments and covariates are low. To overcome this difficulty, we consider approximate balance rather than exact balance, which has been demonstrated to work well in practice in both low- and high-dimensional settings (Athey et al., 2018; Wang & Zubizarreta, 2020; Zubizarreta, 2015b). Specifically, we now choose weights that approximately satisfy the conditions in Equation (3) while minimizing the Kullback–Leibler divergence:

$$
\min \sum_{i=1}^{n} w_i \log(w_i)
$$

s.t.

$$
\left| \frac{1}{n} \sum_{i=1}^{n} w_i \alpha_{K1,l}(T_i) \beta_{K2,\bar{l}}(X_i) - \left( \frac{1}{n} \sum_{i=1}^{n} \alpha_{K1,l}(T_i) \right) \left( \frac{1}{n} \sum_{i=1}^{n} \beta_{K2,\bar{l}}(X_i) \right) \right| \leq \lambda_{l,\bar{l}},
$$

(4)

where $\alpha_{K1,l}(T_i)$ and $\beta_{K2,\bar{l}}(X_i)$ denote the $l$th and $\bar{l}$th components of $\alpha_{K1}(T_i)$ and $\beta_{K2}(X_i)$, respectively. Let $m_K(T_i, X_i) = \text{vec}(\frac{1}{n} \alpha_{K1}(T_i) \beta_{K2}(X_i)')$ and $b_K = \text{vec}(\frac{1}{n} \bar{\alpha}_{K1} \bar{\beta}_{K2}')$ denote two column vectors with dimension $K$, where $K = K1K2$, the $l$th and $\bar{l}$th components of $\bar{\alpha}_{K1}$ and $\bar{\beta}_{K2}$ are defined as

$$
\bar{\alpha}_{K1,l} = \frac{1}{n} \sum_{i=1}^{n} \alpha_{K1,l}(T_i) \text{ and } \bar{\beta}_{K2,l} = \frac{1}{n} \sum_{i=1}^{n} \beta_{K2,l}(X_i).
$$

(5)

Therefore, condition (4) is equivalent to

$$
\min \sum_{i=1}^{n} w_i \log(w_i)
$$
\[
\sum_{i=1}^{n} w_i m_{K,k}(T_i, X_i) - nb_{K,k} \leq \lambda_k, \quad k = 1, \ldots, K. \number{6}
\]

Let both \(\alpha_{K1}(T)\) and \(\beta_{K2}(X)\) contain the constant 1, then we have \(\frac{1}{n} \sum_{i=1}^{n} w_i = 1\) if \(\lambda_1 = 0\) and the primal problem (6) is equivalent to

\[
\min_w \sum_{i=1}^{n} w_i \log(w_i)
\]

s.t.

\[
\sum_{i=1}^{n} w_i (m_{K,k}(T_i, X_i) - b_{K,k}) \leq \lambda_k, \quad k = 1, \ldots, K. \number{7}
\]

The primal problem (7) is difficult to solve numerically, and its dual problem is considered here, which can be solved by numerically efficient algorithms. Theorem 1 provides the dual formulation of problem (7) as an unconstrained problem and helps us to articulate the role of approximate balancing constraints.

**Theorem 1.** The dual of Problem (7) is equivalent to the following unconstrained problem:

\[
\min_{\delta} \sum_{i=1}^{n} \{\exp\{-(m_{K}(T_i, X_i) - b_{K})' \delta - 1\}\} + |\delta|' \lambda,
\]

where \(\delta_{K} = (\delta_{1}, \ldots, \delta_{K})'\) is the vector of dual variables associated with the \(K\) balancing constraints, \(\lambda = (\lambda_1, \ldots, \lambda_K)'\), \(b_{K} = (b_{K,1}, \ldots, b_{K,K})\), and \(m_{K}(T_i, X_i) = (m_1(T_i, X_i), \ldots, m_K(T_i, X_i))\). Besides, the primal solution \(\hat{w}_i\) satisfies

\[
\hat{w}_i = \exp\{-(m_{K}(T_i, X_i) - b_{K})' \hat{\delta} - 1\}, \quad i = 1, \ldots, n,
\]

where \(\hat{\delta}\) is the solution to the dual optimization problem (8).

The proof of Theorem 1 is presented in Appendix A.2. The key to this result is the form of the constraints (7), which allows us to eliminate the positivity constraints on the dual variables. In practice, the tuning parameters \(\lambda\) are selected by the algorithm proposed by Wang and Zubizarreta (2020).

### 3.3 CAUSAL EFFECT ESTIMATION

In this subsection, both parametric and nonparametric approaches are developed to estimate the causal effect function. A weighted optimization estimation is defined under the parametric framework, and B-splines are used to approximate the causal effect function under the nonparametric framework.

#### 3.3.1 Parametric method

The causal effect function is parameterized as \(s(t; \theta)\), we assume that it has a unique solution \(\theta^* \in R^J\) (with \(J \in \mathbb{N}\)) defined as

\[
\theta^* = \arg\min_{\theta} \int_{T} E[Y(t) - s(t; \theta)]^2 f_T(t) dt.
\]

The difficulty in solving Equation (10) is that the potential outcome \(Y(t)\) is not observed for all \(t\). Hence, Proposition 1 is proposed to connect the potential outcome with the observed outcome.
**Proposition 1.** Under Assumption 1, it can be shown that

\[
\mathbb{E}[w(Y - s(T; \theta))^2] = \int_T \mathbb{E}[Y(t) - s(t; \theta)]^2 f_T(t) dt. \tag{11}
\]

The proof of Proposition 1 can be found in Appendix A.1. Note that \(Y(t)\) on the right-hand side of Equation (11) represents the potential outcome and \(Y\) on the left-hand side represents the observed outcome. Proposition 1 indicates that by inserting \(w\) on the left-hand side of Equation (11), one can represent the objective function with the potential outcome by that with the observed outcome. Therefore, the true value \(\theta^*\) is also a solution for the weighted optimization problem:

\[
\theta^* = \arg\min_\theta \mathbb{E}[w(Y - s(T; \theta))^2]. \tag{12}
\]

This result implies that the true value \(\theta^*\) can be identified from the observational data. One can obtain the estimator based on the sample, which is

\[
\hat{\theta} = \arg\min_\theta \sum_{i=1}^{n} \hat{w}_i (Y_i - s(T_i; \theta))^2. \tag{13}
\]

### 3.3.2 Nonparametric method

Suppose \(\mathbb{E}(Y(t)) = s(t)\). In a similar manner to the proof of Proposition 1, it can be shown that

\[
\mathbb{E}(wY | T = t) = \mathbb{E}(Y(t)).
\]

In this paper, B-splines are used to approximate \(s(t)\). For notational convenience, assume that without loss of generality, all B-splines of order \(r\) are defined on an extended partition associated with a uniform partition of \(m\) knots. Following Schumaker (2007), we denote the B-spline basis functions on the \(i\)th component of \(T\) as \(B_j(t_k)(j = 1, \ldots, M; k = 1, \ldots, p; M = m + r)\). Furthermore, define

\[
B_{k_1, \ldots, k_p}(t) = \prod_{j=1}^{p} B_{k_j}(t_j), \forall 1 \leq k_1, \ldots, k_p \leq M. \tag{14}
\]

Let \(B(t)\) be the \(Q \equiv M^p\)-dimensional vector consisting of all product functions of the form (14) and \(Z_n = (B(T_1), \ldots, B(T_n))\), then the B-spline estimation of \(s(t)\) is determined by

\[
\hat{s}(t) = B(t)' \hat{\beta}, \tag{15}
\]

where

\[
\hat{\beta} = (Z_n'Z_n)^{-1}Z_n'\hat{W}Y, \quad \hat{W} = \text{diag}(\hat{w}_1, \ldots, \hat{w}_n).
\]

### 4 LARGE SAMPLE PROPERTIES

To establish the large sample properties of the proposed estimators in Section 3.3, we show the consistency and limiting distribution of the optimization estimator \(\hat{\theta}\) under the parametric framework and the convergence rate of the nonparametric estimator \(\hat{s}(t)\). The following assumptions are made:

**Assumption 4.**

(i) The minimizer \(\delta^* = \arg\min_{\delta \in \Theta_1} \mathbb{E}[\exp[-M(T_1, X_i)' \delta - 1]]\) is unique, where \(\Theta_1\) is the parameter space of \(\delta\) and \(M(T_1, X_i) = m(T_1, X_i) - b_k\).
(ii) $\delta^* \in \text{int}(\Theta_1)$, where $\Theta_1$ is a compact set and $\text{int}(\cdot)$ denotes the interior of a set.

(iii) There exists a constant $c_0$ such that $0 < c_0 < 1$, and $c_0 \leq \exp(-z - 1) \leq 1 - c_0$ for any $z = M(t, x)^T \delta$ with $\delta \in \text{int}(\Theta_1)$.

Besides, there exist constants $c_1 < c_2 < 0$ such that $c_1 \leq -\exp(-z - 1) \leq c_2 < 0$ in some neighborhood of $z^* = M(t, x)^T \delta^*$.

(iv) There exists a constant $C$ such that $\sup_{(t, x)} | M(t, x) | \leq CK^{1/2}$ and $E[M(T_i, X_i)M(T_i, X_i)^T] \leq C$. Besides, $K = o(n)$.

(v) There exist constants $\gamma > 1$ and $\delta^*$ such that $\sup_{(t, x)} | \{\exp(-w - 1)\}^{-1} - M(t, x) \delta^* | = O(K^{-\gamma})$.

(vi) $| \lambda | = O_p \{K^{1/2}(\log K)/n + K^{1-\gamma} \}$.

Assumption 5.

(i) The parameter space $\Theta$ is a compact set, and the true parameter $\theta_0$ is in the interior of $\Theta$.

(ii) $(Y - s(T; \theta))^2$ is continuous in $\theta$, $E[\sup_{\theta} (Y - s(T; \theta))^2] < \infty$, and $\sup_{\theta} E[(Y - s(T; \theta))^4] < \infty$.

Assumption 6.

(i) $s(t; \theta)$ is twice continuously differentiable in $\theta \in \Theta$, and let $h(t; \theta) \equiv \nabla_\theta s(t; \theta)$.

(ii) $E[w(Y - s(T; \theta)) h(T; \theta)]$ is differentiable with respect to $\theta$, and $U \equiv -\nabla_\theta E[w(Y - s(T; \theta)) h(T; \theta)] |_{\theta = \theta^*}$ is nonsingular.

(iii) $E[\sup_{|\theta - \theta^*| < \delta_1} | s(T; \theta_1) - s(T; \theta) |^2]^{1/2} < a \cdot \delta_1^b$ for any $\theta \in \Theta$ and any small $\delta_1 > 0$.

Assumption 4 is the adaptation of Assumption 1 proposed by Wang and Zubizarreta (2020), which ensures the estimated weights are consistent with the stabilized weight. Assumption 5(i) is a commonly used assumption in nonparametric regression. Assumption 5(ii) is an envelope condition applicable to the uniform law of large numbers. Assumptions 6(i) and 6(ii) impose sufficient regularity conditions on the causal effect function and its derivative function. Assumption 6(iii) is a stochastic equicontinuity condition, which is needed for establishing weak convergence (Andrews, 1994). Based on these assumptions, the following theorems are established.

Theorem 2. Let $\hat{\theta}$ denotes the solution to Problem (7), and $\hat{w} = \exp\{-M(t, x)^T \delta - 1\}$, then under Assumptions 1–4,

(i) $\sup_{(t, x)} | \hat{w} - w | = O_p \{K(\log K)/n + K^{1-\gamma} \}$.

(ii) $\frac{1}{n} \sum_{i=1}^n | \hat{w}_i - w_i |^2 = O_p \{K(\log K)^2/n^2 + K^{1-2\gamma} \}$.

Theorem 3.

(i) Under Assumptions 1–5, $\| \hat{\theta} - \theta^* \| \to_p 0$.

(ii) Under Assumptions 1–6, $\sqrt{n}(\hat{\theta} - \theta^*) \to_d N(0, V)$, where

$$V = 4U^{-1} \cdot E[w^2(Y - s(T; \theta^*))^2 h(T; \theta^*) h(T; \theta^*)^T] \cdot U^{-1}.$$ 

Under the nonparametric framework, Theorem 2 is established to obtain the convergence rate of the estimate $\hat{s}(t) = B(t)^T \hat{\beta}$.

Theorem 4. Suppose $\sup_{t \in T} | s(t) - B(t)^T \beta^* | = O(Q^{-\delta})$ holds for some $\delta$ and $\beta^* \in R^Q$, and Assumptions 1–5 hold, then

$$\int | \hat{s}(t) - s(t) |^2 f_T(t) dt = O_p \left( Q^{-2\delta} + \frac{Q}{n} + K^2(\log K)^2/n^2 + K^{2-2\gamma} \right)$$

$$\sup_t | \hat{s}(t) - s(t) | = O_p \{Q^{-\delta} + \sqrt{\frac{Q}{n}} + K(\log K)/n + K^{1-\gamma} \}.$$ 

The proofs of Theorems 2–4 can be found in Appendix A.3.
5 | VARIANCE ESTIMATION AND CONFIDENCE INTERVAL

5.1 | Variance estimation

To conduct statistical inference, a consistent estimator of the covariance matrix \( V \) is needed, which can be obtained by replacing \( w \) and \( \theta^* \) by their consistent estimators. Specifically, one can write \( U \) as

\[
U = -\nabla_{\theta} \mathbb{E}\{w(Y - s(T;\theta))h(T;\theta)\} \bigg|_{\theta=\theta^*} = -\mathbb{E}\{w\nabla_{\theta}[(Y - s(T;\theta))h(T;\theta)]\} \bigg|_{\theta=\theta^*}
\]

\[
= -\mathbb{E}[w[-h(T;\theta)h(T;\theta) + (Y - s(T;\theta))\nabla_{\theta} h(T;\theta)]] \bigg|_{\theta=\theta^*}
\]

\[
= \mathbb{E}[w[h(T;\theta^*)h(T;\theta^*)' - (Y - s(T;\theta^*))\nabla_{\theta} h(T;\theta^*)]].
\]

Hence,

\[
\hat{U} = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i [h(T_i;\hat{\theta})h(T_i;\hat{\theta})' - (Y_i - s(T_i;\hat{\theta}))\nabla_{\theta} h(T_i;\hat{\theta})].
\]

Then we have

\[
\hat{V} = 4\hat{U}^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i^2 (Y_i - s(T_i;\hat{\theta}))^2 h(T_i;\hat{\theta})h(T_i;\hat{\theta}').
\]

According to Theorems 3(i) and 2, we have \( ||\hat{\theta} - \theta^*||_p \to 0 \) and \( \sup |\hat{w} - w| = o_p(1) \), which implies that \( \hat{V} \) is consistent.

5.2 | Confidence interval

According to the asymptotic normality of \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)' \), one can construct the 95% confidence interval of \( \hat{\theta}_j \) as

\[
[\hat{\theta}_j - 1.96 \cdot \hat{S}E_j, \hat{\theta}_j + 1.96 \cdot \hat{S}E_j], \forall j = 1, \ldots, J,
\]

where \( \hat{S}E_j = \hat{V}_{jj}/\sqrt{n} \) is the standard error of \( \hat{\theta}_j \), \( \hat{V}_{jj} \) is the \((j, j)\)-element of \( \hat{V} \).

Alternatively, one can also construct the confidence interval using the bootstrap method. Suppose the \( b \)th bootstrap sample \( \{Y_i^{(b)}, T_i^{(b)}, X_i^{(b)}\}, i = 1, \ldots, n; b = 1, \ldots, B \) is sampled with replacement from the original sample \( \{Y_i, T_i, X_i\}, i = 1, \ldots, n \) with uniform distribution. Denote \( \hat{\theta}^{(b)} \) as the estimator of \( \theta^* \) based on the \( b \)th bootstrap sample. Arrange the \( B \) bootstrap estimators in order from smallest to largest, then the 95% confidence interval can be defined as

\[
[\hat{\theta}^{(0.025B)}, \hat{\theta}^{(0.975B)}], \forall j = 1, \ldots, J.
\]

The difference between the two methods is that the first method (16) relies on the asymptotic normality result while the second method (17) does not.

6 | SIMULATION

To examine the properties of the proposed estimators under finite samples, simulation studies are performed under different data settings. The main motivation in designing the simulation settings is to compare the proposed method with four other methods when the treatment assignment model and the outcome model are linear and nonlinear in various ways.
6.1 Assessment criteria of covariate balance and effect estimation

Assume that the treatments follow the multiple multivariate linear regression model:

\[ T_i = B'X_i + \varepsilon_i, \quad i = 1, \ldots, n, \quad (18) \]

where \( T_i \) and \( X_i \) denote the \( p \)-dimensional treatments and \( q \)-dimensional covariates, respectively. \( B_{q \times p} = (\beta_1 : \beta_2 : \ldots : \beta_q) \) represents the coefficient matrix, and \( \varepsilon_i \sim N_q(0, \Sigma) \). The maximum likelihood estimates (MLEs) of \( B \) are \( \hat{B}_{q \times p} = (X'X)^{-1}X'T \), where \( T = (T'_1, \ldots, T'_n)' \), \( X = (X'_1, \ldots, X'_n)' \).

To assess the performance of the covariate balance, the following hypothesis test is considered:

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0. \]

The likelihood ratio statistic is

\[ \Lambda = \left| \frac{SSE}{SSE + SSH} \right|^\frac{1}{2} \]

where \( SSE = T'T - \hat{B}'X'T \) and \( SSH = \hat{B}'X'T - n\hat{T}'\hat{T} \) denote the residual sum of squares and the predicted sum of squares, respectively.

When \( n \) is large,

\[ -2\log(\Lambda) \rightarrow \chi^2_{q \times p}. \]

Hence, it is harder to reject the null hypothesis when \( -2\log(\Lambda) \) is closer to zero, which implies that the covariate balance performs well.

Next, we focus on the outcome model, which is linear and nonlinear in the treatments, respectively.

\[ Y_i = r(X_i) + T_i\beta + \varepsilon_i, \quad (19) \]

\[ Y_i = r(X_i) + s(T_i) + \varepsilon_i, \quad (20) \]

where \( \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n. \)

For the linear outcome model, the bias and root mean square error (RMSE) of the coefficients are used to assess the performance of the different methods in various cases. For the nonparametric outcome model, we use

\[ \text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{s}(t)_i - s(t)_i)^2} \]

as the assessment criteria.

6.2 Data-generating process

The data-generating process is described in this subsection. For the treatment assignment model, two major cases are considered, which are T-2d-L and T-2d-NL, respectively, where “T” stands for treatment, “2d” stands for two dimensions, and “L” and “NL” stands for linear and nonlinear in the covariates. For the outcome model, linear and nonlinear cases are studied, which are considered for each treatment assignment model. For notational convenience, “Specification YK” is used to denote the \( K \)th outcome model.

Since the covariates are shared across all scenarios, their data-generating process is first described. Specifically, we independently draw five covariates from a multivariate normal distribution with mean 0, variance 1, and covariance 0.2,
that is,

\[ \mathbf{X} = (X_1, \ldots, X_5)' \sim N_5(\mu, \Sigma) \text{ with } \mu = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 1 & 0.2 & \cdots & 0.2 \\ 0.2 & 1 & 0.2 & \cdots & 0.2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0.2 & 0.2 & \cdots & 1 \end{pmatrix}_{5 \times 5}. \]

In the first simulation setting, assume that both the treatment assignment model “T-2d-L” and the outcome model are linear in the covariates, with true data-generating process given by

\[ \mathbf{T} - 2d - L : T_i = B_1' \mathbf{X}_i + \epsilon_i, \quad (21) \]

Specification Y1:

\[ Y_i = T_{i1} + T_{i2} + X_{i1} + 0.1X_{i2} + 0.1X_{i5} + \epsilon_i. \]

Specification Y2:

\[ Y_i = T_{i1} + T_{i2} + 0.2T_{i1}T_{i2} + 0.2T_{i1}X_{i1} + X_{i1} + 0.1X_{i2} + 0.1X_{i5} + \epsilon_i. \]

Specification Y3:

\[ Y_i = T_{i1} + T_{i2} + (T_{i1} - T_{i2})^2 + X_{i1} + 0.1X_{i2} + 0.1X_{i5} + \epsilon_i, \]

where

\[ \mathbf{T}_i = (T_{i1}, T_{i2})', \quad B_1 = \begin{bmatrix} 1 & 1 \\ 0 & 0.2 \\ 0.2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}_{5 \times 2}, \quad \epsilon_i \sim N(0, 2^2), \]

\[ \epsilon_i \sim N_2(\mu, \Sigma) \text{ with } \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 3 & 0.8 \\ 0.8 & 3 \end{pmatrix}. \]

In the second simulation setting, consider the case when the treatment assignment model is misspecified and the outcome model remains the same as in the first simulation setting, that is, the treatment model is nonlinear in the covariates, denoted as “T-2d-NL”:

\[ \mathbf{T} - 2d - NL : T_i = B_1' \mathbf{X}_i + B_2'(\mathbf{X}_i \ast \mathbf{X}_i) + \epsilon_i, \quad (22) \]

where \( B_1 \) and \( \epsilon_i \) are the same as in Equation (20), \( \ast \) represents the Hadamard product of two matrices, and the corresponding element of \( \mathbf{X} \ast \mathbf{X} \) is \( (\mathbf{X} \ast \mathbf{X})_{ij} = (x_{ij}x_{ij}) \), \( B_2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}_{5 \times 2} \).

In the third simulation setting, we consider misspecification of the outcome model by including the nonlinear covariate term in the outcome model. In this case, the generating process for the treatment model remains the same as in the first simulation setting. Specifically, the misspecified outcome model is defined as

Specification Y4:

\[ Y_i = T_{i1} + T_{i2} + (X_{i1} + 1)^3 + 0.1X_{i2} + 0.1X_{i5} + \epsilon_i. \]
Specification $Y_5$:

$$Y_i = T_{i1} + T_{i2} + 0.2T_{i1}T_{i2} + 0.2T_{i1}X_{i1} + (X_{i1} + 1)^3 + 0.1X_{i2} + 0.1X_{i5} + \varepsilon_i.$$ 

Specification $Y_6$:

$$Y_i = T_{i1} + T_{i2} + (T_{i1} - T_{i2})^2 + (X_{i1} + 1)^3 + 0.1X_{i2} + 0.1X_{i5} + \varepsilon_i,$$

where $\varepsilon_i \sim N(0,2^2)$.

The last simulation setting considers the misspecification of both the treatment assignment model and the outcome model, which are defined in the second simulation setting and the third simulation setting, respectively.

To compute the estimated weights, two sieve basis functions $\alpha_{K1}(T)$ and $\beta_{K2}(X)$ need to be specified. For $\alpha_{K1}(T)$, $(1, T' \cdot T' \cdot T'')', (1, T' \cdot T' \cdot (T * T'))', (1, T' \cdot (T * T * T'))'$ are considered. For $\beta_{K2}(X)$, $(1, X' \cdot X'), (1, X' \cdot (X * X'))', (1, X' \cdot (X * X * X'))'$ are considered. Since there are nine pairs of $(K_1, K_2)$ in total, the optimal value of $(K_1^*, K_2^*)$ is defined as

$$(K_1^*, K_2^*) = \arg\min_{(K_1, K_2)} \sum_{i=1}^{n} \tilde{w}_i(Y_i - s(T_i; \hat{\theta}))^2.$$ 

### 6.3 Simulation results

In this subsection, simulation results are compared between the proposed method (EBMT), the regression covariate adjustment method (RCAM), entropy balancing for univariate treatment method (EBUT), the multivariate generalized propensity score method (mvGPS) (Williams & Crespi, 2020), and subclassification method (Subclass) (Imai & Dyk, 2004), where RCAM refers to estimating the causal effect function by adjusting for all covariates in the linear regression outcome model, EBUT refers to estimating the causal effect of the multivariate treatments by the entropy-balancing method for each univariate continuous treatment separately, mvGPS refers to the method that first estimates the stabilized weight by a parametric method assuming that the propensity function follows a multivariate normal distribution and then estimates the causal effect function by conducting a weighted analysis based on the estimated stabilized weight. Subclass refers to the method introduced in Section 2, which is proposed by Imai and Dyk (2004). For the Subclass method, we consider $2 \times 2$ subclassification for simplicity.

Specifically, the performance of covariate balancing is compared among EBMT, mvGPS, and unweighted methods, since RCAM and Subclass methods do not balance covariates and EBUT does not fit multiple multivariate regression so that the balancing effect cannot be measured. When there is no interaction effect between treatments, the estimation accuracy of all five methods is compared. But when there is the interaction between treatments, only four methods (EBMT, RCAM, mvGPS, Subclass) are compared since the method EBUT cannot be applied to this case. In addition, the 95% coverage probability (CP) and the average width (AW) of the confidence interval are considered. These are obtained using the bootstrap method with $B = 500$ iterations based on Equation (17) for each effect estimation. For each numerical setting, 1000 independent simulation experiments are run.

Figure 1 shows the results for the covariate balancing of the linear two-dimensional treatment assignment model (top) and nonlinear two-dimensional treatment assignment model (bottom) under different settings. The statistics $-2\log(\Lambda)$ produced by the EBMT method are almost zero, while those of the mvGPS and unweighted methods are far away from zero. These indicate that the EBMT method balances the covariates well and is robust when the treatment assignment model is linear and nonlinear in the covariates.

Table 1 shows the results of the causal effect estimation for the linear outcome model without interaction between the two-dimensional treatments under different simulation settings. As can be seen, the results for the main effect are similar to those in Table 1, that is, EBMT performs the best...
in most cases. For the interaction effect, the mean biases are slightly larger in most cases. Similarly, all methods fail when both the two models are misspecified.

Table 3 shows the results of the 95% coverage probability and the average width of the bootstrap confidence interval of the causal effect estimation. It can be seen that the 95% coverage probability of the EBMT method is nearly identical to 0.95, and it is larger than the other methods under different scenarios except for the case that both the treatment assignment model and outcome model are misspecified. For all methods, the average width of the confidence interval shrinks as the sample size grows.

Table 4 shows the mean RMSE of the estimated causal effect function for the nonparametric outcome model under different scenarios. Since only the EBMT and mvGPS methods can be applied to such a problem, the results of those two methods are compared. It can be seen that the EBMT method enjoys a better performance compared with the mvGPS method when both the treatment assignment model and the outcome model are correctly specified or either one of the two models is correctly specified. Both methods fail when both the two models are misspecified.

The simulation results can be summarized as follows. (i) In the case that both the treatment assignment model and outcome model are correctly specified, EBMT has the smallest mean bias of the main effect and slightly larger RMSE than RCAM. (ii) In the case that the treatment assignment model is correctly specified and the outcome model is misspecified, EBMT has the smallest mean bias and RMSE of all effects. (iii) In the case that both the treatment assignment model and outcome model are misspecified, all methods fail. (iv) In the case of the nonparametric outcome model, EBMT has the smallest mean RMSE.

7 | APPLICATION

We now apply the EBMT method to the motivating example described in Section 2. The goal is to analyze the impact of the duration and frequency of smoking on medical expenditures.
The pretreatment covariates $X$ includes age at the time of the survey, gender, race, marital status, education level, census region, poverty status, and seat belt usage, which are the covariates used in Imai et al. (2004). The treatments we are interested in are the duration and frequency of smoking, and the outcome is log(Total medical expenditure). Similar to the setting of Imai et al. (2004), we conduct a complete-case analysis by discarding all missing data, yielding a sample of 7847 smokers. The two sieve basis functions $\alpha_{K1}(T)$ and $\alpha_{K2}(X)$ are specified as follows. For $\alpha_{K1}(T)$, $(1, T')'$, $(1, T', (T * T)')'$, and $(1, T', (T * T)')'$, $(T * T * T)'$ are considered. For $\alpha_{K2}(X)$, $(1, X)'$, $(1, X', (X * X)'$, and $(1, X', (X * X)')'$ are considered. Since there are nine pairs of $(K_1, K_2)$ in total, we use the criteria in Section 6.2 to select the optimal value of $(K_1^*, K_2^*)$ and the results show that the optimal sieve basis functions are $(1, T')'$ and $(1, X')'$, respectively.

Before estimating the causal effect, we first examine the covariate balancing of EBMT and mvGPS methods by conducting a multiple multivariate regression. The statistic $-2\log(\Lambda)$ defined in Section 6.1 is $7.77e^{-8}$ for the EBMT method, 2.27

### Table 1: Mean bias and RMSE of the coefficients in the linear outcome model without interaction between the two-dimensional treatments.

| Method   | Both $E(T \mid X)$ and $E(Y \mid X)$ correctly specified | $\beta_1^*(\beta_2^* = 1)$ | $n = 500$ | $n = 1000$ | $\beta_2^*(\beta_1^* = 1)$ | $n = 500$ | $n = 1000$ |
|----------|----------------------------------------------------------|-----------------------------|-----------|-----------|----------------------------|-----------|-----------|
|          |                                                          | $\hat{\beta}_1$             | Bias/RMSE |           | $\hat{\beta}_2$             | Bias/RMSE |           |
|          |                                                          | $\hat{\beta}_2$             |           | Bias/RMSE |                           |           | Bias/RMSE |
| EBMT     | $-0.004/0.083$                                           | $0.005/0.051$               |           | $0.007/0.186$ | $0.006/0.117$               |           |           |
| RCAM     | $-0.006/0.058$                                           | $-0.005/0.043$              |           | $0.011/0.104$ | $0.010/0.081$               |           |           |
| EBUT     | $-0.003/0.084$                                           | $-0.002/0.058$              |           | $0.008/0.206$ | $0.007/0.133$               |           |           |
| mvGPS    | $-0.012/0.132$                                           | $0.006/0.082$               |           | $0.151/0.319$ | $0.035/0.171$               |           |           |
| Subclass | $-0.005/0.078$                                           | $0.002/0.048$               |           | $0.009/0.118$ | $0.008/0.095$               |           |           |
|          | $E(T \mid X)$ correctly specified, $E(Y \mid X)$ mispecified | $\hat{\beta}_1^*(\beta_2^* = 1)$ | $n = 500$ | $n = 1000$ | $\hat{\beta}_2^*(\beta_1^* = 0.8)$ | $n = 500$ | $n = 1000$ |
|          |                                                          | $\hat{\beta}_1$             | Bias/RMSE |           | $\hat{\beta}_2$             | Bias/RMSE |           |
|          |                                                          | $\hat{\beta}_2$             |           | Bias/RMSE |                           |           | Bias/RMSE |
| EBMT     | $-0.009/0.125$                                           | $-0.005/0.087$              |           | $0.019/0.276$ | $0.028/0.178$               |           |           |
| RCAM     | $-0.001/0.051$                                           | $0.007/0.067$               |           | $0.031/0.333$ | $-0.047/0.194$              |           |           |
| EBUT     | $0.014/0.189$                                            | $0.022/0.144$               |           | $0.023/0.284$ | $0.034/0.180$               |           |           |
| mvGPS    | $-0.075/0.767$                                           | $-0.016/0.591$              |           | $1.035/1.408$ | $0.762/1.051$               |           |           |
| Subclass | $-0.010/0.162$                                           | $0.006/0.113$               |           | $0.025/0.353$ | $-0.031/0.191$              |           |           |
|          | $E(T \mid X)$ misspecified, $E(Y \mid X)$ correctly specified | $\hat{\beta}_1^*(\beta_2^* = 0.8)$ | $n = 500$ | $n = 1000$ | $\hat{\beta}_2^*(\beta_1^* = 0.8)$ | $n = 500$ | $n = 1000$ |
|          |                                                          | $\hat{\beta}_1$             | Bias/RMSE |           | $\hat{\beta}_2$             | Bias/RMSE |           |
|          |                                                          | $\hat{\beta}_2$             |           | Bias/RMSE |                           |           | Bias/RMSE |
| EBMT     | $-0.004/0.071$                                           | $0.006/0.052$               |           | $0.008/0.131$ | $-0.005/0.086$              |           |           |
| RCAM     | $-0.007/0.067$                                           | $0.011/0.048$               |           | $0.014/0.090$ | $-0.009/0.083$              |           |           |
| EBUT     | $-0.017/0.114$                                           | $0.010/0.086$               |           | $0.012/0.147$ | $-0.008/0.105$              |           |           |
| mvGPS    | $-0.069/0.642$                                           | $-0.019/0.347$              |           | $-0.380/0.813$ | $-0.183/0.514$              |           |           |
| Subclass | $-0.008/0.116$                                           | $0.010/0.082$               |           | $0.013/0.187$ | $-0.009/0.122$              |           |           |

Abbreviations: EBMT, entropy balancing for multivariate treatment; mvGPS, multivariate generalized propensity score method; RCAM, regression covariate adjustment method; RSME, root mean square error.
### TABLE 2
Mean bias and RMSE of the coefficients in the linear outcome model with interaction between the two-dimensional treatments.

| Method   | Both $E(T \mid X)$ and $E(Y \mid X)$ correctly specified | $\hat{\beta}_1 (\beta^*_1 = 1)$ | $n = 500$ | Bias/RMSE | $n = 1000$ | Bias/RMSE | $\hat{\beta}_2 (\beta^*_2 = 0.8)$ | $n = 500$ | Bias/RMSE | $n = 1000$ | Bias/RMSE | $\hat{\beta}_{12} (\beta^*_1 \beta^*_2 = 0.2)$ | $n = 500$ | Bias/RMSE | $n = 1000$ | Bias/RMSE |
|----------|----------------------------------------------------------|----------------------------------|-----------|-----------|-----------|-----------|-----------------------------------|-----------|-----------|-----------|-----------|-----------------------------------|-----------|-----------|-----------|-----------|
|          |                                                          | $n = 500$                        | $n = 1000$|
| EBMT     |                                                          | 0.003/0.092                     | 0.005/0.063|
| RCAM     |                                                          | 0.009/0.067                     | 0.007/0.043|
| mvGPS    |                                                          | 0.011/0.139                     | 0.013/0.123|
| Subclass |                                                          | 0.008/0.071                     | -0.006/0.041|
|          |                                                          | $n = 500$                        | $n = 1000$|
| EBMT     |                                                          | 0.009/0.112                     | 0.006/0.086|
| RCAM     |                                                          | 0.024/0.151                     | 0.012/0.106|
| mvGPS    |                                                          | 0.176/0.796                     | -0.067/0.430|
| Subclass |                                                          | 0.022/0.153                     | 0.009/0.109|
|          |                                                          | $n = 500$                        | $n = 1000$|
| EBMT     |                                                          | 0.004/0.031                     | 0.002/0.029|
| RCAM     |                                                          | 0.012/0.024                     | 0.008/0.017|
| mvGPS    |                                                          | 0.006/0.045                     | 0.004/0.031|
| Subclass |                                                          | 0.008/0.033                     | 0.007/0.030|
|          |                                                          | $n = 500$                        | $n = 1000$|
| EBMT     |                                                          | 0.095/0.178                     | 0.155/0.209|
| RCAM     |                                                          | -0.265/0.278                    | -0.254/0.268|
| mvGPS    |                                                          | 0.418/0.949                     | 0.573/0.997|
| Subclass |                                                          | -0.261/0.301                    | -0.253/0.287|

Abbreviations: EBMT, entropy balancing for multivariate treatment; mvGPS, multivariate generalized propensity score method; RCAM, regression covariate adjustment method; RMSE, root mean square error.

for the mvGPS method, and 2.31 for the unweighted method, which shows that EBMT balances covariates very well while mvGPS does not balance well.

### 7.1 Parametric approach

The causal effect function is estimated by the parametric method assuming a linear model of the outcome. The bootstrap method is used to obtain the standard errors and confidence intervals for the parameter estimates. For simplicity, we consider $2 \times 2$ subclassification for the Subclass method.

Table 5 shows the estimated causal effects of the duration and frequency of smoking on medical expenditure as well as their standard errors and confidence intervals. All methods indicate that the duration of smoking has no significant impact on medical expenditure since their confidence intervals all cover zero while most methods agree that the frequency of smoking increases medical expenditure significantly except mvGPS. Besides, the width of the confidence interval of the estimated causal effect based on the EBMT method is the smallest, which implies the estimation accuracy of the EBMT method is the largest. Compared with univariate analysis using a single variable `packyear`, the analysis of the bivariate
TABLE 3  Confidence interval estimation of the coefficients in the linear outcome model without interaction between the two-dimensional treatments.

| Method   | Both \(E(T \mid X)\) and \(E(Y \mid X)\) correctly specified | \(E(T \mid X)\) correctly specified, \(E(Y \mid X)\) misspecified | \(E(T \mid X)\) misspecified, \(E(Y \mid X)\) correctly specified | Both \(E(T \mid X)\) and \(E(Y \mid X)\) misspecified |
|----------|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|---------------------------------------------------------------|
|          | \(\hat{\beta}_1 (\hat{\gamma}_1^* = 1)\) \(\hat{\beta}_2 (\hat{\gamma}_2^* = 1)\) | \(\hat{\beta}_1 (\hat{\gamma}_1^* = 1)\) \(\hat{\beta}_2 (\hat{\gamma}_2^* = 0.8)\) | \(\hat{\beta}_1 (\hat{\gamma}_1^* = 1)\) \(\hat{\beta}_2 (\hat{\gamma}_2^* = 0.8)\) | \(\hat{\beta}_1 (\hat{\gamma}_1^* = 1)\) \(\hat{\beta}_2 (\hat{\gamma}_2^* = 0.8)\) |
|          | \(n = 500\) \(95\%\ CP/AW\) \(n = 1000\) \(95\%\ CP/AW\) | \(n = 500\) \(95\%\ CP/AW\) \(n = 1000\) \(95\%\ CP/AW\) | \(n = 500\) \(95\%\ CP/AW\) \(n = 1000\) \(95\%\ CP/AW\) | \(n = 500\) \(95\%\ CP/AW\) \(n = 1000\) \(95\%\ CP/AW\) |
| EBMT     | 0.951/0.327 0.955/0.223 | 0.945/0.683 0.954/0.479 | 0.914/0.385 0.938/0.314 | 0.028/0.757 0.031/0.617 |
| RCAM     | 0.926/0.222 0.942/0.162 | 0.934/0.438 0.952/0.318 | 0.906/0.383 0.913/0.321 | 0.013/0.775 0.019/0.538 |
| EIBUT    | 0.944/0.306 0.953/0.226 | 0.877/0.734 0.914/0.527 | 0.835/2.156 0.837/2.156 | 0.037/3.303 0.043/2.714 |
| mvGPS    | 0.948/0.434 0.954/0.340 | 0.758/0.723 0.812/0.598 | 0.943/0.518 0.953/0.405 | 0.004/1.207 0.007/0.896 |
| Subclass | 0.934/0.310 0.946/0.177 | 0.943/0.518 0.953/0.405 | 0.935/1.463 0.942/0.952 | 0.921/0.371 0.938/0.342 |

Abbreviations: EBMT, entropy balancing for multivariate treatment; mvGPS, multivariate generalized propensity score method; RCAM, regression covariate adjustment method.

treatments provides more information that the significant effect of \(\text{pack year}\) attributes mostly to the frequency of smoking rather than to its duration.

7.2  Nonparametric approach

In practice, the linearity assumption of the outcome model may be violated; hence, the nonparametric method is also applied to explore the causal relationship between the frequency and duration of smoking and the medical expenditures, which is displayed in Figure 2.

Figure 2a shows that there is an overall upward trend, which indicates that the frequency and duration of smoking generally have a positive effect on medical costs. Observe that there is a slight downward trend at the edges when both
TABLE 4 Mean RMSE comparison of the nonparametric outcome model of the two-dimensional treatments.

| Method   | $n = 500$                              | $n = 1000$                              |
|----------|----------------------------------------|----------------------------------------|
|          | Both $E(T \mid X)$ and $E(Y \mid X)$ correctly specified |                                      |
| EBMT     | 0.944                                  | 0.792                                  |
| mvGPS    | 1.650                                  | 1.289                                  |
|          | $E(T \mid X)$ correctly specified, $E(Y \mid X)$ misspecified |                                      |
| EBMT     | 3.144                                  | 2.967                                  |
| mvGPS    | 4.655                                  | 4.468                                  |
|          | $E(T \mid X)$ misspecified, $E(Y \mid X)$ correctly specified |                                      |
| EBMT     | 0.952                                  | 0.821                                  |
| mvGPS    | 1.835                                  | 2.361                                  |
|          | Both $E(T \mid X)$ and $E(Y \mid X)$ misspecified |                                      |
| EBMT     | 6.326                                  | 5.885                                  |
| mvGPS    | 5.583                                  | 4.581                                  |

Abbreviations: EBMT, entropy balancing for multivariate treatment; mvGPS, multivariate generalized propensity score method; RMSE, root mean square error.

TABLE 5 Causal effect estimation of the duration and frequency of smoking on medical expenditure.

| Method   | Duration | Frequency |
|----------|----------|-----------|
|          | Estimate | SE        | 95% CI    | Estimate | SE        | 95% CI    |
| EBMT     | 0.003    | 0.002     | (−0.004,0.003) | 0.007    | 0.002     | (0.003,0.010) |
| RCAM     | −0.001   | 0.004     | (−0.004,0.010) | 0.006    | 0.003     | (0.001,0.013) |
| EBUT     | 0.002    | 0.003     | (−0.005,0.008) | 0.007    | 0.002     | (0.002,0.012) |
| mvGPS    | −0.074   | 0.143     | (−0.420,0.099) | 0.018    | 0.046     | (−0.075,0.098) |
| Subclass | 0.001    | 0.003     | (−0.006,0.005) | 0.007    | 0.002     | (0.004,0.011) |

Abbreviations: EBMT, entropy balancing for multivariate treatment; mvGPS, multivariate generalized propensity score method; RCAM, regression covariate adjustment method; SE, standard error.

frequency and duration of smoking are very small, which may be because those who smoke really little might suffer from other diseases which cause a little higher medical costs than those who smoke a certain amount. However, these are just our guesses since there are no covariates describing the health status of the subjects available. Compared with the parametric method, Figure 2a provides more information about the relationship between the duration and frequency of smoking and medical expenditures. Figure 2b shows that there are many observations near the corner of low duration and low frequency, which indicates that the nonlinear relationship near this corner is reasonable.

8 CONCLUSION AND DISCUSSION

In this study, we extend the one-dimensional entropy-balancing method to multidimensional treatments. In addition, parametric and nonparametric methods are developed to estimate the causal effect and their theoretical properties are provided. The simulation results show that the proposed method balances covariates well and produces a smaller mean bias compared with other methods. In the real-data analysis, the EBMT method is applied to investigate the causal relationship between the duration and frequency of smoking on medical expenditure. The results from the parametric method indicate that the frequency of smoking increases medical costs significantly. The results from the nonparametric method indicate that there is a short-term downward trend and a long-term upward trend when both the duration and frequency of smoking increase.

In this paper, we mainly consider the causal effect function $E(Y(t))$ as the estimand for continuous treatment as it is general. The average treatment effect or average partial effect can also be considered, but they can be easily estimated based on the estimates of causal effect function $E(Y(t + \Delta t) - Y(t))$ (Viet, 2008). Indeed, the causal effect function provides a complete description of the causal effect, rather than a summary measure. Besides, parametric and nonparametric methods are developed to estimate the causal effect function. Since parametric methods are easy to imple-
ment and the sample size requirement is not high, one may use them when reasonable assumptions can be made about the true model. Nonparametric methods enjoy higher flexibility and less model assumptions but have higher requirements for the sample size. One may choose to use them according to the real situations.

Future works include (1) considering treatments that involve other types of complex data, such as longitudinal data or functional data; (2) extending this method to the case of multivariate outcome variables.

ACKNOWLEDGMENTS
The work was supported by the National Natural Science Foundation of China (project number: 11771146, 11831008), the 111 Project (B14019) and the Program of Shanghai Subject Chief Scientist (14XD1401600).

CONFLICT OF INTEREST STATEMENT
The authors declare no conflicts of interest.
DATA AVAILABILITY STATEMENT
The R code and the simulated data therein are shared. The application data are made publicly available in Imai et al. (2004), which can be found at DOI: https://doi.org/10.1198/016214504000001187.

OPEN RESEARCH BADGES
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID
Yingchun Zhou https://orcid.org/0000-0003-1217-6989

REFERENCES
Ai, C., Linton, O., Motegi, K., & Zhang, Z. (2021). A unified framework for efficient estimation of general treatment models. Quantitative Economics, 12(3), 779–816.
Andrews, D. W. (1994). Empirical process methods in econometrics. Handbook of Econometrics, 4, 2247–2294.
Athey, S., Imbens, G. W., & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society Series B, 80(Part 4), 597–623.
Chan, K. C. G., Yam, S. C. P., & Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. Journal of the Royal Statistical Society: Series B, 78(3), 673.
Dasgupta, T., Pillai, N. S., & Rubin, D. B. (2015). Causal inference from 2k factorial designs by using potential outcomes. Journal of the Royal Statistical Society: Series B, 77(4), 727–753.
Dong, Y., Lee, Y.-Y., & Gou, M. (2021). Regression discontinuity designs with a continuous treatment. Journal of the American Statistical Association. Advanced online publication. https://doi.org/10.1080/01621459.2021.1923509
Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. The Annals of Applied Statistics, 12(1), 156–177.
Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1), 25–46.
Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. Political Analysis, 22(1), 1–30.
Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In A. Gelman & X.-L. Meng (Eds.), Applied Bayesian modeling and causal inference from incomplete-data perspectives: An essential journey with Donald Rubin’s statistical family (pp. 73–84). Wiley.
Hsu, Y.-C., Lai, T.-C., & Lieli, R. P. (2022). Counterfactual treatment effects: Estimation and inference. Journal of Business & Economic Statistics, 40(1), 240–255.
Huber, M., Hsu, Y.-C., Lee, Y.-Y., & Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. Journal of Applied Econometrics, 35(7), 814–840.
Imai, K., & Dyk, D. V. (2004). Causal inference with general treatment regimes. Journal of the American Statistical Association, 99(467), 854–866.
Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B, 76(1), 243–263.
Imai, K., & Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association, 99(467), 854–866.
Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87(3), 706–710.
Johnson, E., Dominici, F., Griswold, M., & Zeger, S. L. (2003). Disease cases and their medical costs attributable to smoking: An analysis of the National Medical Expenditure Survey. Journal of Econometrics, 112(1), 135–151.
Larsen, M. D. (1999). An analysis of survey data on smoking using propensity scores. Sankhýyā: The Indian Journal of Statistics, Series B, 61, 91–105.
Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of Econometrics, 4, 2111–2245.
Robbins, M. W., Griffin, B. A., Shih, R. A., & Slaughter, M. E. (2020). Robust estimation of the causal effect of time-varying neighborhood factors on health outcomes. Statistics in Medicine, 39(5), 544–561.
Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 550–560.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55.
Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, 79(387), 516–524.
Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39(1), 33–38.
Rubin, D. B. (2000). Statistical issues in the estimation of the causal effects of smoking due to the conduct of the tobacco industry. In J. L. Gastwirth (Ed.), *Statistical science in the courtroom* (pp. 321–351). Springer.

Schumaker, L. (2007). *Spline functions: Basic theory*. Cambridge University Press.

Tseng, P., & Bertsekas, D. P. (1987). Relaxation methods for problems with strictly convex separable costs and linear constraints. *Mathematical Programming, 38*(3), 303–321.

Tseng, P., & Bertsekas, D. P. (1991). Relaxation methods for problems with strictly convex costs and linear constraints. *Mathematics of Operations Research, 16*(3), 462–481.

Tropp, J. A., et al. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning, 8*(1–2), 1–230.

Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Vegetabile, B. G., Gillen, D. L., & Stern, H. S. (2020). Optimally balanced Gaussian process propensity scores for estimating treatment effects. *Journal of the Royal Statistical Society: Series A, 183*(1), 355–377.

Viet, C. N. (2008). A note on estimation of the average treatment effect and average partial effect in nonlinear models. *Economics Bulletin, 15*, 1–13.

Wang, Y., & Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika, 107*(1), 93–105.

Williams, J. R., & Crespi, C. M. (2020). *Causal inference for multiple continuous exposures via the multivariate generalized propensity score*. arXiv. https://doi.org/10.48550/arXiv.2008.13767

Xiong, C., Yu, M., & Shao, J. (2017). Treatment recommendation and parameter estimation under single-index contrast function. *Statistical Theory and Related Fields, 1*(2), 171–181.

Yiu, S., & Su, L. (2018). Covariate association eliminating weights: a unified weighting framework for causal effect estimation. *Biometrika, 105*(3), 709–722.

Zeger, S. L., Wyant, T., Miller, L. S., & Samet, J. (2000). Statistical testimony on damages in Minnesota v. tobacco industry. In J. L. Gastwirth (Ed.), *Statistical science in the courtroom* (pp. 303–320). Springer, New York.

Zhe, Q., & Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference, 5*(1), 20160010.

Zhu, Y., Coffman, D. L., & Ghosh, D. (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference, 3*(1), 25–40.

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

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

**Supporting Information**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Chen, J., & Zhou, Y. (2023). Causal effect estimation for multivariate continuous treatments. *Biometrical Journal, 65*, 2200122. https://doi.org/10.1002/bimj.202200122

**Appendix**

A.1 | Proof of Proposition 1

Using the law of total expectation and Assumption 1, we can deduce that

\[
\mathbb{E}[w(Y - s(T; \theta))^2] = \mathbb{E}\left[ \frac{f(T)}{f(T | X)} (Y - s(T; \theta))^2 \right]
= \mathbb{E}(\mathbb{E}\left[ \frac{f(T)}{f(T | X)} (Y - s(T; \theta))^2 \right] | T = t, X = x)
= \mathbb{E}(\frac{f(t)}{f(t | x)} \mathbb{E}((Y - s(T; \theta))^2 | T = t, X = x))
= \int_{T \times X} \frac{f(t)}{f(t | x)} \mathbb{E}((Y(T) - s(T; \theta))^2 | T = t, X = x) f(t | x) dt dx
\]
\[
\int_{T \times X} E[(Y(T) - s(T; \theta))^2 \mid T = t, X = x] f(t) f(x) dt dx \\
= \int_{T \times X} E[(Y(t) - s(t; \theta))^2 \mid X = x] f(t) f(x) dt dx \\
= \int_{T \times X} E[(Y(t) - s(t; \theta))^2 f(t) dt \quad \text{(using Assumption 1)}.
\]

Hence, we complete the proof of Proposition 1.

A.2 \quad \textbf{Proof of Theorem 1}

To prove Theorem 1, we first rewrite problem (6) in matrix notation,

\[
\begin{align*}
\text{minimize}_{w} & \sum_{i=1}^{n} w_i \ln(w_i) \\
\text{subject to} & \quad Q_{2K \times n} w_n \leq d_{2K \times 1},
\end{align*}
\]

where

\[
\begin{align*}
w_n &= (w_i)_{n \times 1}, \\
A_{K \times n} &= \begin{pmatrix} m_{K,1}(T_1, X_1) - b_{K,1} & m_{K,1}(T_2, X_2) - b_{K,1} & \cdots & m_{K,1}(T_n, X_n) - b_{K,1} \\
\vdots & \vdots & \ddots & \vdots \\
m_{K,K}(T_1, X_1) - b_{K,K} & m_{K,K}(T_2, X_2) - b_{K,K} & \cdots & m_{K,K}(T_n, X_n) - b_{K,K} \end{pmatrix}, \\
Q_{2K \times n} &= \begin{pmatrix} A_{K \times n} \\
-A_{K \times n} \end{pmatrix}, \\
\lambda_{K \times 1} &= (\lambda_1, \ldots, \lambda_K) \quad \text{and} \\
d_{2K \times 1} &= \begin{pmatrix} \lambda_{K \times 1} \\
\lambda_{K \times 1} \end{pmatrix}.
\end{align*}
\]

The problem is now in the form of Tseng and Bertsekas (1987, 1991).

The dual of this problem is

\[
\begin{align*}
\text{maximize} & \quad \delta g(\delta) \\
\text{subject to} & \quad \delta \geq 0,
\end{align*}
\]

where

\[
g(\delta) = -\sum_{i=1}^{n} h^*(Q_i \delta) - \langle \delta, d \rangle \quad \text{and}
\]

\[
h^*(t) = \sup_{w_i} \{-tw_i - w_i \ln(w_i)\} = -tw_i^* - w_i^* \ln(w_i^*),
\]

where \(w_i^*\) satisfies that \(-t - \ln(w_i^*) - 1 = 0\), hence \(w_i^* = e^{-t-1}\). Furthermore,

\[
h^*(t) = -t \ln(e^{-t-1}) - (-t - 1)e^{-t-1} = e^{-t-1}.
\]

Let \(\rho(t) = e^{-t-1}\), then \(\rho'(t) = e^{-t-1} = h^*(t)\) and \(w_i^* = \rho'(t)\).

Hence, the dual formulation becomes

\[
\begin{align*}
\text{minimize}_{\delta} & \quad l(\delta) \\
\text{subject to} & \quad \delta \geq 0,
\end{align*}
\]

where

\[
l(\delta) = \sum_{i=1}^{n} \rho'(Q_i \delta) + \delta' d = \sum_{i=1}^{n} e^{-Q_i \delta - 1} + \delta' d.
\]
Next, we write \( \delta_{2K} = \begin{pmatrix} \delta_{+,K} \\ \delta_{-,K} \end{pmatrix} \), then

\[
I(\delta) = \sum_{i=1}^{n} \rho'(A_i'\delta_+ - A_i'\delta_-) + \delta_+\lambda + \delta_-\lambda
\]

\[
= \sum_{i=1}^{n} \rho'(A_i'(\delta_+ - \delta_-)) + (\delta_+ + \delta_-)\lambda.
\]

Suppose the optimizer is \( \delta_{2K}^* = \begin{pmatrix} \delta_{+,K}^* \\ \delta_{-,K}^* \end{pmatrix} \), then it can be shown that \( \delta_{+,k}^* \cdot \delta_{-,k}^* = 0, \ k = 1, \ldots, K \). Theorem 1 then follows by rewriting \( \delta = \delta_+ - \delta_- \) and \( |\delta| = \delta_+ + \delta_- \); hence, the proof of Theorem 1 is completed.

A.3 | Large sample properties

A.3.1 | Proof of Theorem 2

To prove Theorem 2, we first prove the following lemma:

**Lemma A1.** Under Assumption 1–4, we have

\[
\| \hat{\delta} - \delta^* \| = O_p(K^{1/2}(\log K)/n + K^{1/2-r}).
\]

**Proof.** Let

\[
A_j = m_K(T_j, X_j) - b_K,
\]

\[
G(\delta) = \frac{1}{n} \sum_{i=1}^{n} \rho'(A_i'\delta) + \frac{1}{n} |\delta|\lambda.
\]

It can be shown that \( G(\cdot) \) is convex in \( \delta \). To show a minimizer \( \Delta^* \) of \( G(\delta^* + \Delta) \) exists in \( C = \{ \Delta \in \mathbb{R}^K : \|\Delta\|_2 \leq CK^{1/2}(\log K)/n + K^{1/2-r} \} \) for some constant \( C \), it suffices to show that

\[
E\{\inf_{\Delta \in C} G(\delta^* + \Delta) - G(\delta^*) > 0\} \to 1 \text{ as } n \to \infty
\]

(A1)

by the continuity of \( G(\cdot) \).

To show Equation (21), we use the mean value theorem:

\[
G(\delta^* + \Delta) - G(\delta^*) \geq \Delta \cdot \frac{1}{n} \sum_{i=1}^{n} \rho''(A_i'\delta^*)A_i + \frac{1}{2n} \Delta' \cdot \left\{ \sum_{i=1}^{n} \rho'''(A_i'\delta_1)A_iA_i \right\} \cdot \Delta - \frac{1}{n} \| \Delta \|_2 \| \lambda \|_2
\]

\[
\geq - \| \Delta \|_2 \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \rho''(A_i'\delta^*)A_i \right\|_2 + \frac{1}{2n} \Delta' \cdot \left\{ \sum_{i=1}^{n} \rho'''(A_i'\delta_1)A_iA_i \right\} \cdot \Delta - \frac{1}{n} \| \Delta \|_2 \| \lambda \|_2
\]

\[
\geq - \| \Delta \|_2 \cdot \left\| \frac{1}{n} \sum_{i=1}^{n} \rho''(A_i'\delta^*)A_i \right\|_2 - \| \Delta \|_2 \| \lambda \|_2.
\]

The first inequality follows the fact that \( |\delta^* + \Delta| - |\delta^*| \geq - |\Delta| \). The second inequality follows from the Cauchy–Schwarz inequality. The third inequality is due to the positivity of \( \frac{1}{2n} \Delta' \cdot \left\{ \sum_{i=1}^{n} \rho'''(A_i'\delta_1)A_iA_i \right\} \cdot \Delta \).
Next, we note that
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \rho''(A_i' \delta^*) A_i \right\|_2 \leq \left\| \frac{1}{n} \sum_{i=1}^{n} (\rho''(A_i' \delta^*) + w_i - 1) A_i \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^{n} (1 - w_i) A_i \right\|_2 \\
\leq \frac{1}{n} \sum_{i=1}^{n} \| A_i \|_2 O(K^{-r}) + \left\| \frac{1}{n} \sum_{i=1}^{n} (1 - w_i) A_i \right\|_2.
\]

The first inequality is due to the triangle inequality. The second inequality is due to Assumption 4(v).

Since
\[
\left\| \frac{1}{n} (1 - w_i) A_i \right\|_2 \leq \frac{1}{n} \| (1 - w_i) \|_2 \| A_i \|_2 \\
\leq \frac{1}{n} \sup | w_i - 1 | CK^{1/2} \\
\leq C_1 K^{1/2} / n.
\]

The first inequality is due to the Cauchy–Schwarz inequality. The second and third inequalities are due to Assumptions 4(iii) and 4(iv), respectively.

Besides, we have
\[
\left\| \sum_{i=1}^{n} E\left\{ \frac{1}{n^2} (1 - w_i)^2 A_i A_i' \right\} \right\|_2 \leq \frac{1}{n} \sup (1 - w_i)^2 \| E(A_i A_i') \|_2 \\
\leq \frac{C_2}{n}.
\]

Combing Equations (A1) and (A2) and the Bernstein’s inequality (Tropp et al. (2015)), one can show that
\[
P \left( \left\| \frac{1}{n} \sum_{i=1}^{n} (1 - w_i) A_i \right\|_2 \geq t \right) \leq K \exp \left( - \frac{t^2}{2} \frac{C_2}{n} + C_1 \frac{K^{1/2} \log K}{n} \right).
\]

The right-hand side becomes zero as \(K \to \infty\) when
\[
\frac{t^2}{2} \frac{C_2}{n} + C_1 \frac{K^{1/2} \log K}{n} \geq \log K.
\]

Hence, it suffices when \(t = O_p(K^{1/2} (\log K)/n)\), which implies that
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} (1 - w_i) A_i \right\|_2 = O_p(K^{1/2} (\log K)/n).
\]

According to Assumption 4(iv), we can obtain that
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} A_i \right\|_2 O(K^{-r}) \leq CK^{1/2 - r}.
\]
Combing Equations (A3) and (A4) and Assumption 4(vi), we have

\[ G(\delta^* + \Delta) - G(\delta^*) \geq -\| \Delta \|_2 \cdot O_p(K^{1/2}(\log K)/n + K^{1/2-r}) + \frac{1}{2} \| \Delta \|_2^2 \| \lambda \|_2 \]

\[ \geq 0 \]

for \( \Delta = \frac{nK^{1/2}\log K}{n} + K^{1/2-r} \) with a large enough constant \( C > 0 \); hence, Equation (21) is proved. Therefore, the proof of Lemma A1 is completed.

Now we prove Theorem 2. Since

\[
\sup_{(t,x)} | \hat{w} - w^* | = \sup_{(t,x)} | \rho'(M(t,x)'\hat{\delta}) - \rho'(M(t,x)'\delta^*) | 
\leq |M(t,x)'\delta_1| \times \sup_{(t,x)} | M(t,x)'(\hat{\delta} - \delta^*) | 
\leq O_p(1) \cdot \| \hat{\delta} - \delta^* \| \cdot \sup_{(t,x)} || M(t,x) || 
\leq O_p(1) \cdot O_p(K^{1/2}(\log K)/n + K^{1/2-r}) \cdot O(K^{1/2}) 
= O_p(K\log K/n + K^{1-r}) 
= o_p(1),
\]

where \( \delta_1 \) lies between \( \hat{\delta} \) and \( \delta^* \).

By the mean value theorem, we can deduce that

\[
\int | \hat{w} - w^* |^2 dF(t,x) 
\leq \sup_{(t,x)} | \rho''(M(t,x)'\delta_1) |^2 \times \int | M(t,x)'(\hat{\delta} - \delta^*) |^2 dF(t,x) 
\leq O_p(1) \cdot \int | M(t,x)'(\hat{\delta} - \delta^*) |^2 dF(t,x).
\]

Since

\[
\int | M(t,x)'(\hat{\delta} - \delta^*) |^2 dF(t,x) 
= \int M(t,x)'(\hat{\delta} - \delta^*)(\hat{\delta} - \delta^*)'M(t,x)dF(t,x) 
= \text{tr}((\hat{\delta} - \delta^*)(\hat{\delta} - \delta^*)' \int M(t,x)M(t,x)'dF(t,x)) 
\leq C \text{tr}((\hat{\delta} - \delta^*)(\hat{\delta} - \delta^*)') 
= C \| \hat{\delta} - \delta^* \|^2 
= O_p(K(\log K)^2/n^2 + K^{1-2r}).
\]

Then we have

\[
\int | \hat{w} - w^* |^2 dF(t,x) = O_p(K(\log K)^2/n^2 + K^{1-2r}).
\]
Furthermore, one can show that
\[
\frac{1}{n} \sum_{i=1}^{n} | M(t, x)'(\hat{\delta} - \delta^*) |^2 - \int | M(t, x)'(\delta - \delta^*) |^2 dF(t, x)
= O_p((K/n)^{1/2}) || \hat{\delta} - \delta^* ||^2
= o_p(1),
\]
where the last equality holds in light of Assumption 4(iv). Hence,
\[
\frac{1}{n} \sum_{i=1}^{n} | \hat{w}_i - w_i^* |^2
\leq \sup_{t, x} | \rho''(M(t, x)' \delta_1) |^2 \cdot \frac{1}{n} \sum_{i=1}^{n} | M(t, x)'(\delta - \delta^*) |^2
\leq O_p(1) \int | M(t, x)'(\hat{\delta} - \delta^*) |^2 dF(t, x) + o_p(1)
= O_p(K (\log K)^2 / n^2 + K^{1-2r}).
\]

Therefore, the proof of Theorem 2 is completed.

### A.3.2 | Proof of Theorem 3

We first show the conclusion of Theorem 1.

Since \( \hat{\delta} \) (as an estimator of \( \theta^* \)) is a unique minimizer of \( \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i (Y_i - s(T_i; \theta))^2 \) (regarding \( \mathbb{E}[w(Y - s(T; \theta))^2] \), according to the theory of M-estimation (Van der Vaart, 2000, Theorem 5.7), if
\[
\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i (Y_i - s(T_i; \theta))^2 - \mathbb{E}[w(Y - s(T; \theta))^2] \right| \rightarrow_p 0,
\]
then \( \hat{\delta} \rightarrow_p \theta^* \). Note that
\[
\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i (Y_i - s(T_i; \theta))^2 - \mathbb{E}[w(Y - s(T; \theta))^2] \right| \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i - w_i) (Y_i - s(T_i; \theta))^2 \right|,
\]
\[
+ \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - s(T_i; \theta))^2 - \mathbb{E}[w(Y - s(T; \theta))^2] \right|.
\]

We first show that \( \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i - w_i) (Y_i - s(T_i; \theta))^2 \right| = o_p(1) \). Using the Cauchy–Schwarz inequality and the fact that \( \hat{w} \rightarrow L^2 \) \( w \), we have
\[
\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i - w_i) (Y_i - s(T_i; \theta))^2 \right| \leq \left\{ \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i - w_i)^2 \right\}^{1/2} \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s(T_i; \theta))^2 \right\}^{1/2}
\leq o_p(1) \left( \sup_{\theta \in \Theta} \mathbb{E}[w(Y - s(T; \theta))^2] + o_p(1) \right)^{1/2}
= o_p(1).
\]
Thereafter, under Assumption 6, we can conclude that \( \sup_{\theta \in \Theta} | \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - s(T_i; \theta))^2 - \mathbb{E}[w(Y - s(T; \theta))^2] | \) is also \( o_p(1) \) (Newey & McFadden, 1994, Lemma 2.4). Hence, we complete the proof for Theorem 3(i). Next, we give the proof of Theorem 3(ii).

Define

\[
\hat{\theta}^* = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - s(T_i; \theta))^2.
\]

Assume that \( \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - s(T_i; \hat{\theta}^*))h(T_i; \hat{\theta}^*)) = o_p(n^{-1/2}) \) holds with probability to one as \( n \to \infty \).

By Assumption 5 and the uniform law of large numbers, one can get that

\[
\frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - s(T_i; \theta))^2 \to \mathbb{E}[w(Y - s(T; \theta))^2] \text{ in probability uniformly over } \theta,
\]

which implies \( ||\hat{\theta}^* - \theta^*|| \to_p 0 \). Let

\[
r(\theta) = 2\mathbb{E}\{w(Y - s(T; \theta))h(T; \theta)\},
\]

which is a differentiable function in \( \theta \) and \( r(\theta^*) = 0 \). By the mean value theorem, we have

\[
\sqrt{n}r(\hat{\theta}^*) - \nabla_{\theta}r(\zeta) \cdot \sqrt{n}(\hat{\theta}^* - \theta^*) = \sqrt{n}r(\theta^*) = 0,
\]

where \( \zeta \) lies on the line joining \( \hat{\theta}^* \) and \( \theta^* \). Since \( \nabla_{\theta}r(\theta) \) is continuous at \( \theta^* \) and \( ||\hat{\theta}^* - \theta^*|| \to_p 0 \), then

\[
\sqrt{n}(\hat{\theta}^* - \theta^*) = \nabla_{\theta}r(\theta^*)^{-1} \cdot \sqrt{n}r(\hat{\theta}^*) + o_p(1).
\]

Define the empirical process

\[
G_n(\theta) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \{w_i (Y_i - s(T_i; \theta))h(T_i; \theta) - \mathbb{E}[w(Y - s(T; \theta))h(T; \theta)]\}.
\]

Then, we have

\[
\sqrt{n}(\hat{\theta}^* - \theta^*) = \nabla_{\theta}r(\theta^*)^{-1} \cdot \left\{ \sqrt{n}r(\hat{\theta}^*) - \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \{w_i (Y_i - s(T_i; \hat{\theta}^*))h(T_i; \hat{\theta}^*) + \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \{w_i (Y_i - s(T_i; \theta^*))h(T_i; \theta^*)\}\right\}
\]

\[
= -\nabla_{\theta}r(\theta^*)^{-1} \cdot G_n(\hat{\theta}^*) + o_p(1)
\]

\[
= U^{-1} \cdot \{G_n(\hat{\theta}^*) - G_n(\theta^*) + G_n(\theta^*)\} + o_p(1).
\]

By Assumptions 5 and 6 and Theorems 4 and 5 of Andrews (1994), we have \( G_n(\hat{\theta}^*) - G_n(\theta^*) \to_p 0 \). Thus,

\[
\sqrt{n}(\hat{\theta}^* - \theta^*) = U^{-1} \frac{2}{\sqrt{n}} \sum_{i=1}^{n} \{w_i (Y_i - s(T_i; \theta^*))h(T_i; \theta^*)\} + o_p(1),
\]
then we can get that the asymptotic variance of $\sqrt{n}(\hat{\theta}^* - \theta^*)$ is $V$. Therefore, $\sqrt{n}(\hat{\theta}^* - \theta^*) \to_d N(0, V)$. Next, we will prove $\hat{\theta} \to_p \hat{\theta}^*$. Since

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \hat{w}_i (Y_i - s(\mathbf{T}_i; \theta))^2 - \frac{1}{n} \sum_{i=1}^{n} w_i (Y_i - s(\mathbf{T}_i; \theta))^2 \right| \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i - w_i) (Y_i - s(\mathbf{T}_i; \theta))^2 \right|$$

$$\leq \left\{ \frac{1}{n} \sum_{i=1}^{n} (\hat{w}_i - w_i)^2 \right\}^{1/2} \sup_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - s(\mathbf{T}_i; \theta))^2 \right\}^{1/2}$$

$$\leq o_p(1) \left[ \sup_{\theta \in \Theta} \mathbb{E}[\hat{w}(Y - s(\mathbf{T}; \theta))^2] + o_p(1) \right]^{1/2} = o_p(1),$$

which implies $\hat{\theta}^* \to_p \hat{\theta}$. Then by Slutsky’s theorem, we can conclude that $\sqrt{n}(\hat{\theta} - \theta^*) \to_d N(0, V)$. Therefore, we have completed the proof of Theorem 3.

A.3.3 | Proof of Theorem 4

To obtain the convergence rate of the estimate $\hat{s}(\mathbf{t}) = B(\mathbf{t})^T \hat{\beta}$, we first obtain the convergence rate of $\hat{\beta}$.

Note that

$$\hat{\beta} - \beta^* = (Z'_n Z_n)^{-1} Z_n' \hat{w} Y - \beta^*$$

$$= (Z'_n Z_n)^{-1} Z_n' (\hat{w} - w) Y + (Z'_n Z_n)^{-1} Z_n' (w Y - E(w Y \mid \mathbf{T})) + (Z'_n Z_n)^{-1} Z_n' (E(w Y \mid \mathbf{T}) - Z_n \beta^*)$$

$$\equiv B_1 + B_2 + B_3.$$ (A7)

First, we compute the convergence order of $B_1$. Let $H = (\hat{w} - w) Y$ and $\hat{\Lambda} = \frac{1}{n} Z'_n Z_n$; then,

$$|| B_1 ||^2 = \| (n \hat{\Lambda})^{-1} Z_n' H \|^2$$

$$= n^{-2} \text{tr}(H' Z_n (\hat{\Lambda})^{-2} Z_n' H)$$

$$= n^{-2} \text{tr}((\hat{\Lambda})^{-1} Z'_n H H' Z_n (\hat{\Lambda})^{-1})$$

$$= n^{-2} \text{tr}((\hat{\Lambda})^{-1/2} Z'_n H H' Z_n (\hat{\Lambda})^{-1/2} (\hat{\Lambda})^{-1})$$

$$\leq \lambda_{\text{max}}((\hat{\Lambda})^{-1}) n^{-2} \text{tr}((\hat{\Lambda})^{-1/2} Z'_n H H' Z_n (\hat{\Lambda})^{-1/2})$$

$$= \lambda_{\text{max}}((\hat{\Lambda})^{-1}) n^{-2} \text{tr}(H H' Z_n (\hat{\Lambda})^{-1} Z'_n)$$

$$= \lambda_{\text{max}}((\hat{\Lambda})^{-1}) n^{-1} \| H \|^2$$

$$= \lambda_{\text{max}}((\hat{\Lambda})^{-1}) n^{-1} Y' (\hat{w} - w)' (\hat{w} - w) Y$$

$$= \lambda_{\text{max}}((\hat{\Lambda})^{-1}) \sup_{(\mathbf{t}, \mathbf{x})} || w - w ||^2 \frac{1}{n} Y' Y$$

$$\leq O_p(1) O_p(K^2 (\log K)^2 / n^2 + K^{2-2r}) O_p(1)$$

$$= O_p(K^2 (\log K)^2 / n^2 + K^{2-2r}),$$ (A8)

where the first inequality follows from $\text{tr}(AB) \leq \lambda_{\text{max}}(B) \text{tr}(A)$ for any symmetric matrix $B$ and positive semidefinite matrix $A$, the second inequality follows from the fact that $Z_n (Z'_n Z_n)^{-1} Z'_n$ is a projection matrix with a maximum eigenvalue of 1, and the last inequality follows from the fact that $| \lambda_{\text{max}}((\hat{\Lambda})^{-1}) | = O_p(1)$, Theorem 2, and $\frac{1}{n} Y' Y = O_p(1)$. 


Thereafter, we compute the convergence order of $B_2$. Let $E = WY - E(WY \mid T)$, then

$$
|| B_2 ||^2 = || (n\hat{\Lambda})^{-1}Z_n'E ||^2 \\
= n^{-2} \text{tr}(E'Z_n(\hat{\Lambda})^{-2}Z_n'E) \\
= n^{-2} \text{tr}((\hat{\Lambda})^{-1}Z_n'EE'Z_n(\hat{\Lambda})^{-1}) \\
= n^{-2} \text{tr}(Z_n'EE'Z_n(\hat{\Lambda})^{-2}) \\
\leq \lambda_{\max}((\hat{\Lambda})^{-2})n^{-2} \text{tr}(Z_n'EE'Z_n) \\
= \lambda_{\max}((\hat{\Lambda})^{-2})n^{-2} || E'Z_n ||^2 \\
= O_p\left(\frac{Q}{n}\right),
$$

(A9)

where the last equality follows the fact that $|\lambda_{\max}((\hat{\Lambda})^{-1})| = O_p(1)$ and $n^{-2} || E'Z_n ||^2 = O_p\left(\frac{Q}{n}\right)$ by Markov’s inequality.

Finally, we compute the convergence order of $B_3$. Let $M = E(WY \mid T) - Z_n\beta^*$, then

$$
|| B_3 ||^2 = || (n\hat{\Lambda})^{-1}Z_n'M ||^2 \\
= n^{-2} \text{tr}(M'Z_n(\hat{\Lambda})^{-2}Z_n'M) \\
= n^{-2} \text{tr}((\hat{\Lambda})^{-1}Z_n'MM'Z_n(\hat{\Lambda})^{-1}) \\
= n^{-2} \text{tr}((\hat{\Lambda})^{-1/2}Z_n'MM'Z_n(\hat{\Lambda})^{-1/2}) \\
\leq \lambda_{\max}((\hat{\Lambda})^{-1})n^{-2} \text{tr}((\hat{\Lambda})^{-1/2}Z_n'MM'Z_n(\hat{\Lambda})^{-1/2}) \\
= \lambda_{\max}((\hat{\Lambda})^{-1})n^{-2} \text{tr}(Z_n'MM'Z_n(\hat{\Lambda})^{-1}) \\
= \lambda_{\max}((\hat{\Lambda})^{-1})n^{-1} \text{tr}(MM'Z_nZ_n^{-1}Z_n') \\
= \lambda_{\max}((\hat{\Lambda})^{-1})n^{-1} \sum_{i=1}^{n} E(w_iY_i \mid T_i) - B(T_i')\beta^*)^2.
$$

(A10)

Since

$$
n^{-1} \sum_{i=1}^{n} E(w_iY_i \mid T_i) - B(T_i')\beta^*)^2 \leq \sup_{t \in \Theta} | E(wY \mid T = t) - B(t)\beta^* |^2 = O_p(Q^{-2\delta}),
$$

we can deduce that $|| B_3 ||^2 \leq O_p(Q^{-2\delta})$.

By combining Equations (29)–(31), we have

$$
|| \hat{\beta} - \beta^* || = O_p\left(\sqrt{\frac{Q}{n}} + Q^{-\alpha} + K(\log K)/n + K^{1-r}\right).
$$
Next, the proof of the convergence of $\hat{s}(t)$ is provided. Let $\Lambda = E[B(t)'B(t)]$, then we have

$$
\int_{\mathcal{R}} |\hat{s}(t) - s(t)|^2 dF_T(t) = \int_{\mathcal{R}} |B(t)'\hat{\beta} - B(t)'\hat{\beta}^* + B(t)'\hat{\beta}^* - s(t)|^2 dF_T(t)
\leq 2(\hat{\beta} - \beta^*)' \int_{\mathcal{R}} [B(t)'B(t)] dF_T(t)(\hat{\beta} - \beta^*) + 2 \int_{\mathcal{R}} |B(t)'\hat{\beta}^* - s(t)|^2 dF_T(t)
\leq 2 \|\hat{\beta} - \beta^*\|^2 \lambda_{max}(\Lambda) + 2 \int_{\mathcal{R}} |B(t)'\beta^* - s(t)|^2 dF_T(t)
= O_p \left(\frac{Q}{n} + Q^{-2\alpha} + K^2(\log K)^2/n^2 + K^{2-2r}\right)
$$

(A11)

and

$$
\sup_{t \in \mathcal{R}} |\hat{s}(t) - s(t)| = \sup_{t \in \mathcal{R}} |B(t)'\hat{\beta} - B(t)'\hat{\beta}^* + B(t)'\hat{\beta}^* - s(t)|
\leq \sup_{t \in \mathcal{R}} \|B(t)\| \|\hat{\beta} - \beta^*\| + \sup_{t \in \mathcal{R}} |B(t)'\hat{\beta}^* - s(t)|
\leq O_p(\sqrt{\frac{Q}{n} + Q^{-\alpha}}) + O_p(Q^{-\alpha})
= O_p(\sqrt{\frac{Q}{n} + Q^{-\alpha} + K(\log K)/n + K^{1-r}}).
$$

(A12)

Therefore, the proof of Theorem 4 is complete.