Econometric Theory, **42**, 2026, 1194–1227.
doi:10.1017/S0266466625100200

AVERAGING ESTIMATORS OF HETEROGENEOUS TREATMENT EFFECTS UNDER ADDITIVE MODELS

Na Li
Kunming University of Science and Technology

Yu Fei
Yunnan University of Finance and Economics

Yuhong Yang
Tsinghua University and Beijing Institute of Mathematical Sciences and Applications

Xinyu Zhang
Chinese Academy of Sciences and University of Science and Technology of China

We consider spline-based additive models for estimation of conditional treatment effects. To handle the uncertainty due to variable selection, we propose a method of model averaging with weights obtained by minimizing a J-fold cross-validation criterion, in which a nearest neighbor matching is used to approximate the unobserved potential outcomes. We show that the proposed method is asymptotically optimal in the sense of achieving the lowest possible squared loss in some settings and assigning all weight to the correctly specified models if such models exist in the candidate set. Moreover, consistency properties of the optimal weights and model averaging estimators are established. A simulation study and an empirical example demonstrate the superiority of the proposed estimator over other methods.

1. INTRODUCTION

Estimating the treatment effect of an intervention, both in randomized experiments and observational studies, is an important goal in a wide variety of application areas ranging from personalized medicine to social sciences. The intervention may be a drug, a training program, a college degree, an advertisement, or anything else that is intended or likely to alter an outcome. For a given intervention, its

The authors are very grateful to the editor, associate editor, and two anonymous referees for their constructive comments and suggestions. X.Z.'s work is supported by the National Key Research and Development Program of China (Grant No. 2023YFA1008704) and the National Natural Science Foundation of China (Grant Nos. 72525001 and 72495124). Y.F.'s work is supported by the National Natural Science Foundation of China (Grant No. 12561051) and Yunnan Province XingDian Talent Support Program (Grant No. YNWR-YLXZ-2018-020). Address correspondence to Yuhong Yang, Yau Mathematical Sciences Center, Tsinghua University, Beijing, China and Beijing Institute of Mathematical Sciences and Applications, Beijing, China, e-mail: yyangsc@tsinghua.edu.cn.

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

effect may differ across different individuals. For example, an advertisement may impact the consumption behaviors of male and female consumers differently at different ages, so the treatment effect may be heterogeneous within the population. Estimating heterogeneous treatment effects has received growing attention. Under the potential outcome framework, conditional averaging treatment effect (CATE) is used to measure heterogeneous treatment effect. Various estimation methods for CATE have been well studied under both parametric and nonparametric frameworks with different types of learning algorithms. Recent references include Imai and Ratkovic (2013), Athey and Imbens (2016), Künzel et al. (2019), Nie and Wager (2021), and Lewbel, Choi, and Zhou (2023) among others. Although many models and methods for estimating CATE have been proposed, the question of which model or method to use in a specific application needs to be answered.

To address this issue, a popular approach is model selection, which picks a "good" model by using data-driven criteria for CATE estimation (Rolling and Yang, 2014; Koch, Vock, and Wolfson, 2018). However, model selection puts too much confidence in a single model, which may substantially increase the estimation risk. Besides, if the goal is to provide an accurate prediction or estimation, using a single selected model may miss some useful information contained in other candidate models. An alternative approach is to apply model averaging, which deals with model selection uncertainty by assembling estimators from multiple candidate models with certain weights. A primary motivation of model averaging is that it typically provides a protection against selecting a relatively poor model and reduces bias and/or variance of the estimator. As a powerful prediction and estimation tool, model averaging methods have attracted much attention in recent decades. A substantial body of model averaging techniques is available in the literature (Hoeting et al., 1999; Yang, 2000; Hansen, 2007; Hansen and Racine, 2012; Liu, 2015; Liu and Kuo, 2016; Rolling, Yang, and Velez, 2019; Liao et al., 2019; Li et al., 2020; Li et al., 2021; Yuan, Wu, and Fang, 2022; Fang, Li, and Xia, 2022). However, few studies have focused on model averaging estimators for heterogeneous treatment effects. Rolling et al. (2019) develop a model combination method for heterogeneous treatment effects, a technique known as treat effect estimation by mixing (TEEM). This method is motivated by the idea of adaptive regression by mixing that can be used for averaging both parametric and nonparametric models. However, it focuses on a worst-case performance viewpoint, and does not address the goal of asymptotical optimality. In fact, for CATE estimation, model averaging estimators with asymptotically optimal weights are scarce. In the case of parametric models, Zhao et al. (2024) firstly propose an optimal model averaging estimator of treatment effects based on linear models. In many applications, the relationships between covariates and outcome are nonlinear. For instance, Huang, Horowitz, and Wei (2010) document a nonlinear relationship between the expression of the gene TRIM32 and those of the 500 genes in the study of Bardet–Biedl syndrome for rats. It is worth pointing out that when nonparametric models are among the candidates, model selection and model averaging for CATE are harder compared to regular regression (Rolling and

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Yang, 2014). To address the difficulty that one cannot observe both treatment and control outcomes of an individual, a good approach is to create a surrogate of the missing outcome by a proper matching of the individuals in the data.

In this article, we develop a novel model averaging estimator for heterogeneous treatment effects by assembling estimators from multiple additive candidate models. A list of candidate models with different specifications of covariates are constructed using an additive nonlinear structure that is a summation of univariate smooth functions. The additive structure is considered here because of its advantages: (i) it retains the interpretable additive form of linear models (Xue, 2009) and (ii) it is flexible as a nonparametric approach, and may circumvent the "curse of dimensionality" (Stone, 1985). To appropriately choose the optimal averaging weights, we adopt a J-fold cross-validation (CV) criterion, in which a pairing method with dimensional reduction is applied to impute the unobserved potential outcomes. Our theoretical results and implementations are all based on the same matching method. Theoretical properties are derived in two practical scenarios: (i) all candidate models are misspecified and (ii) some correct models are included in the candidate set. In the former scenario, we prove that the proposed method is asymptotically optimal in the sense that its squared estimation loss is asymptotically identical to the infeasible best possible model averaging estimator. In the latter scenario, we demonstrate that the J-fold CV asymptotically assigns all weights to the correctly specified models. In addition, the consistency of the optimal weights and the model averaging estimator are established in the latter scenario.

There are major differences between our method and those in the previous work on model averaging for CATE, especially with Zhao et al. (2024). First, while Zhao et al. (2024) impose the structural constraint involving fixed-dimensional linear parametric models, this framework faces notable practical limitations. Specifically, the assumption appears to be rather restrictive given the nonlinear nature of covariate–outcome relationships commonly seen in many applications, and the situation is even worse when the covariate dimensions are not small. We employ additive models to investigate the nonlinear associations between the response and covariates, while allowing the dimension of the covariates to increase with the sample size. Second, the matching methods to approximate the unobserved potential outcomes are inconsistent between theory and application in Zhao et al. (2024). They employ a partition method to construct the potential outcomes due to an independence requirement needed in their theoretical analysis, but a different matching method is adopted both in simulation and empirical application. In contrast, our theoretical and simulation studies are based on the same nearest neighbor matching. With the aforementioned differences, investigating the asymptotic properties of our proposed model averaging estimator is perhaps more challenging than that in Zhao et al. (2024). Third, unlike Zhao et al. (2024), we additionally study the consistency of the model averaging estimator and the properties of weights when correct models are included in the set of candidate models. Fourth, for each candidate model, the same subset of covariates is used

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

to construct the treatment model and control model in Zhao et al. (2024). For more flexibility, the two additive models can be constructed by using different sets of covariates in our article. Finally, although nonparametric model averaging for asymptotic optimal prediction has been well studied (Hu and Zhang, 2023; Liao et al., 2023), it seems that no such work has focused on the rather different goal of treatment effect estimation. Our work makes an effort in establishing a spline-based additive model averaging estimation for the conditional treatment effects.

Our theoretical results are supported by an extensive set of simulation studies. We also illustrate the proposed method with a data set concerning the effect of macroeconomic uncertainty on household spending. In fact, our estimator achieves a smaller out-of-sample estimation error than all the other methods considered in a guided simulation.

The outline of the article is as follows. Section 2 sets up the model framework, and then the model averaging estimation for the treatment effect is introduced in Section 3. Section 4 establishes asymptotic results when the candidate models are misspecified or the candidate models include at least one correct model. Simulation results comparing different methods are presented in Section 5, and we illustrate the methods on a real data set in Section 6. In Section 7, we draw conclusions and provide further discussions on the proposed method. Technical proofs are given in the Supplementary Material.

2. MODEL FRAMEWORK

We formalize our problem in terms of the potential outcomes framework (Rubin, 1974; Splawa-Neyman, Dabrowska, and Speed, 1990). Suppose that we observe an independent and identically distributed random sample $\{\boldsymbol{U}_i, T_i, Y_i\}_{i=1}^n$ with sample size n, where $\boldsymbol{U}_i = (U_{i,1}, \ldots, U_{i,p})'$ denotes per-person features, $T_i \in \{0,1\}$ is the treatment assignment, and Y_i is the observed outcome. Here, $T_i = 1$ represents that individual i is assigned to the treatment group, and $T_i = 0$ indicates the control group. We posit the existence of potential outcomes $\{Y_i(0), Y_i(1)\}$ corresponding to the outcome we would have observed given the treatment assignment $T_i = 0$ or 1, respectively, such that $Y_i = Y_i(T_i)$. Then, we can write the potential outcome model as

$$\begin{aligned} &Y_i = Y_i(1)\mathbb{I}(T_i = 1) + Y_i(0)\mathbb{I}(T_i = 0), \quad 1 \le i \le n, \\ &Y_i(1) = \mu_1(\boldsymbol{U}_i) + \varepsilon_i^{(1)}, \quad \text{and} \quad Y_i(0) = \mu_0(\boldsymbol{U}_i) + \varepsilon_i^{(0)}, \end{aligned} \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, $\mu_t(\boldsymbol{U}_i) = \mathrm{E}(Y_i(t)|\boldsymbol{U}_i)$ for $t \in \{0,1\}$, $\varepsilon_i^{(1)}$ is a random error under treatment, and $\varepsilon_i^{(0)}$ under control. We assume the covariates $\boldsymbol{U}_i$ to be independent and identically distributed from an unknown probability distribution P_U with a compact and convex support $\mathbb{U} \subset \mathbb{R}^p$. The random errors among individuals are assumed to be independent, and the treatment and control error distributions are allowed to differ. Formally, we assume that

$$\mathrm{E}(\varepsilon_i^{(t)}|\boldsymbol{U}_i = \boldsymbol{u}_i) = 0, \quad \varepsilon_i^{(t)} \perp \varepsilon_j^{(t)}, \quad i \neq j,$$

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

where $\perp$ denotes independence, and

$$\mathrm{Var}(\varepsilon_i^{(t)}|\boldsymbol{U}_i = \boldsymbol{u}_i) = \sigma_t^2(\boldsymbol{u}_i),$$

for $t = 0, 1$, and $i, j = 1, 2, \ldots, n$.

Our goal is to estimate the CATE function. For given $\boldsymbol{U} = \boldsymbol{u}$, the CATE is defined as the conditional mean difference in potential outcomes

$$\mathrm{CATE}(\boldsymbol{u}) = \mathrm{E}(Y(1) - Y(0)|\boldsymbol{U} = \boldsymbol{u}). \tag{2}$$

The definition of the CATE function relies on what is called the counterfactual framework or potential outcomes framework (Rubin, 1974; Splawa-Neyman et al., 1990). A major difficulty for CATE($\boldsymbol{u}$) estimation is that although the potential outcomes, $Y(0)$ and $Y(1)$, are hypothesized to exist, only one of them can be observed. Throughout the article, we make the following conditions.

Condition 1 (Unconfounded assignment). Given the covariate $\boldsymbol{U}_i$, the treatment assignment T_i is independent of the potential outcome,

$$T_i \perp \{Y_i(1), Y_i(0)\}|\boldsymbol{U}_i.$$

Condition 1 is certainly satisfied in randomized studies but may also be justified in observational studies if all the variables affecting the treatment assignment and the outcome are measured. Under Condition 1, we have

$$\begin{aligned}\Delta(\boldsymbol{u}) &=: \mathrm{E}(Y|\boldsymbol{U} = \boldsymbol{u}, T = 1) - \mathrm{E}(Y|\boldsymbol{U} = \boldsymbol{u}, T = 0)\\ &= \mathrm{E}(Y(1) - Y(0)|\boldsymbol{U} = \boldsymbol{u})\\ &= \mathrm{CATE}(\boldsymbol{u}).\end{aligned}$$

Then, we can reexpress CATE($\boldsymbol{u}$) as $\Delta(\boldsymbol{u})$, which suggests an approach of estimating CATE: first, estimate the two regression functions $\mu_0(\boldsymbol{u})$ and $\mu_1(\boldsymbol{u})$ separately by using the treated and control samples, and then take their difference. This estimation procedure leads to $\hat{\Delta}(\boldsymbol{u}) = \hat{\mu}_1(\boldsymbol{u}) - \hat{\mu}_0(\boldsymbol{u})$, which is sometimes called "T-learner" with "T" being short for "two." "T-learner" is a popular method to estimate CATE($\boldsymbol{u}$) (Cai et al., 2011; Min et al., 2018; Künzel et al., 2019), and it is also applied in this work.

Condition 2.

$$\eta < e(\boldsymbol{u}) = \mathrm{P}(T = 1|\boldsymbol{U} = \boldsymbol{u}) < 1 - \eta,$$

for some $0 < \eta < 1$.

Condition 2 effectively guarantees that there will be enough treatment and control observations around any $\boldsymbol{u} \in \mathbb{U}$ for large enough n. Furthermore, it implies that the number of treatment observations (n_1), control observations (n_0), and the total sample size (n) are of the same order in probability.

Notation. For any two sequences $a_n, b_n, n = 1, 2, \ldots$, we write $a_n \asymp b_n$, if there are two constants $0 < c_1 < c_2 < \infty$ such that $c_1 \le a_n/b_n \le c_2$ for all sufficiently large n.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

We use $\bar{\lambda}(\boldsymbol{A})$ and $\underline{\lambda}(\boldsymbol{A})$ to denote the maximum and minimum eigenvalues of symmetric matrix $\boldsymbol{A}$, and $\|\boldsymbol{a}\| = (\sum_{i=1}^{m} a_i^2)^{1/2}$ denotes the l_2 norm of any vector $\boldsymbol{a} \in \mathbb{R}^m$, and $\|\boldsymbol{a}\|_\infty = \max_i |a_i|$. In this work, for any matrix $\boldsymbol{A} = (a_{ij})_{i=1,j=1}^{s1,s2}$, $\|\boldsymbol{A}\| = \sqrt{\bar{\lambda}(\boldsymbol{A}'\boldsymbol{A})}$ denotes the spectral norm of a matrix $\boldsymbol{A}$ and $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le s_1} \sum_{j=1}^{s_2} |a_{ij}|$, and $\boldsymbol{e}_i$ is an $n \times 1$ vector with the ith element being 1, and other elements 0.

2.1. Estimation Method

When multiple covariates are available, it may be difficult to estimate unstructured nonparametric multi-dimensional functions, $\mu_1(\boldsymbol{u})$ and $\mu_0(\boldsymbol{u})$. We use two additive models to approximate the two functions. This additive combination of univariate functions is less general than joint multivariate non-parametric modeling but can be more interpretable and easier to fit. For more flexibility, the two additive models can be constructed by using different sets of covariates.

Let $\boldsymbol{\varsigma}^{(1)} = (j_1^{(1)}, j_2^{(1)}, \ldots, j_{p_1}^{(1)})$ and $\boldsymbol{\varsigma}^{(0)} = (j_1^{(0)}, j_2^{(0)}, \ldots, j_{p_0}^{(0)})$ denote the index of variables used in $\mu_1(\boldsymbol{u})$ and $\mu_0(\boldsymbol{u})$, respectively. Let $\boldsymbol{u}^{(1)} = (u_{j_1^{(1)}}, u_{j_2^{(1)}}, \ldots, u_{j_{p_1}^{(1)}})'$ and $\boldsymbol{u}^{(0)} = (u_{j_1^{(0)}}, u_{j_2^{(0)}}, \ldots, u_{j_{p_0}^{(0)}})'$ be the corresponding selection of $\boldsymbol{u}$, respectively. For the ith subject, the corresponding variables are $\boldsymbol{u}_i^{(1)} = (u_{i,j_1^{(1)}}, u_{i,j_2^{(1)}}, \ldots, u_{i,j_{p_1}^{(1)}})'$ and $\boldsymbol{u}_i^{(0)} = (u_{i,j_1^{(0)}}, u_{i,j_2^{(0)}}, \ldots, u_{i,j_{p_0}^{(0)}})'$. We model $\mu_1(\boldsymbol{u})$ and $\mu_0(\boldsymbol{u})$ as an additive combination of univariate functions in $\boldsymbol{u}^{(1)}$ and $\boldsymbol{u}^{(0)}$, respectively. Thus, $\mu_1(\boldsymbol{u})$ and $\mu_0(\boldsymbol{u})$ are approximated as

$$\mu_1(\boldsymbol{u}) \approx \nu_1 + f_{1,1}(u_{j_1^{(1)}}) + f_{2,1}(u_{j_2^{(1)}}) + \cdots + f_{p_1,1}(u_{j_{p_1}^{(1)}}),$$
$$\mu_0(\boldsymbol{u}) \approx \nu_0 + f_{1,0}(u_{j_1^{(0)}}) + f_{2,0}(u_{j_2^{(0)}}) + \cdots + f_{p_0,0}(u_{j_{p_0}^{(0)}}),$$

where ν_0 and ν_1 are unknown constants, and $\{f_{l,0}\}_{l=1}^{p_0}$ and $\{f_{l,1}\}_{l=1}^{p_1}$ are unknown nonparametric functions. Ultimately, the potential outcome model can be written as

$$\begin{aligned} Y(1) &\approx \nu_1 + f_{1,1}(u_{j_1^{(1)}}) + f_{2,1}(u_{j_2^{(1)}}) + \cdots + f_{p_1,1}(u_{j_{p_1}^{(1)}}) + \varepsilon^{(1)}, \\ Y(0) &\approx \nu_0 + f_{1,0}(u_{j_1^{(0)}}) + f_{2,0}(u_{j_2^{(0)}}) + \cdots + f_{p_0,0}(u_{j_{p_0}^{(0)}}) + \varepsilon^{(0)}. \end{aligned} \tag{3}$$

For identifiability, we assume that

$$\mathrm{E}(f_{l,t}) = 0,$$

for any $l = 1, \ldots, p_t$ and $t = 0, 1$.

Using model (3), $\mu_1(\boldsymbol{u})$ and $\mu_0(\boldsymbol{u})$ are estimated separately based on data $\{(\boldsymbol{u}_i, Y_i) : T_i = 1\}$ and $\{(\boldsymbol{u}_i, Y_i) : T_i = 0\}$, respectively, by one of the two frequently-used fitting methods: kernel regression (Buja, Hastie, and Tibshirani, 1989; Ruppert and Wand, 1994; Opsomer and Ruppert, 1997) or spline regression (Stone, 1985; Buja et al., 1989; Ravikumar et al., 2009; Huang et al., 2010). Compared to spline regression, the kernel regression method is slightly more computationally

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

intensive, especially when facing a large number of covariates. In this article, the B-spline estimation method is used. Suppose that each component of $\boldsymbol{U}_i$ takes values in $[a,b]$, where $a<b$ are finite numbers. Let $a=\xi_0<\xi_1<\cdots<\xi_{\tilde{\Gamma}_n-1}<\xi_{\tilde{\Gamma}_n}=b$ be a partition of $[a,b]$ into $\tilde{\Gamma}_n$ subintervals $I_{nl}=[\xi_l,\xi_{l+1})$, $l=0,1,\ldots,\tilde{\Gamma}_n-1$, where the number of subintervals $\tilde{\Gamma}_n=O(n^{\varrho})$ with $0<\varrho<0.5$, and the length of the largest subinterval $\max_{0\le k\le\tilde{\Gamma}_n-1}|\xi_{t+1}-\xi_t|=O(n^{-\varrho})$. Let $\mathcal{S}_n$ be the space of polynomial splines of degree $\gamma\ge 1$ consisting of functions φ satisfying:

(i) the restriction of φ to the interval I_{nl} is a polynomial of degree γ for $l=0,1,\ldots,\tilde{\Gamma}_n-1$;

(ii) for $\gamma>2$ and $0\le\gamma_1\le\gamma-2$, φ is γ_1 times continuously differentiable on $[a,b]$.

For $\mathcal{S}_{n_t}$, considering a normalized B-spline basis $\phi_k, 1\le k\le m_{n_t}$, where $m_{n_t}\equiv\tilde{\Gamma}_{n_t}+\gamma$ (Schumaker, 1981), for any $f^*_{l,t}\in\mathcal{S}_{n_t}$, we can write

$$f^*_{l,t}(u)=\sum_{k=1}^{m_{n_t}}\beta^{(t)}_{lk}\phi_k(u),\qquad 1\le l\le p_t,\qquad t=0,1.$$

Under suitable smoothness assumptions for the functional components in model (3), $f_{l,t}$, for $1\le l\le p_t$ and $t=0,1$, can be approximated by a function in $\mathcal{S}_{n_t}$.

Let $\boldsymbol{Y}_1(n_1\times 1)=(Y_{11},\ldots,Y_{n_1 1})'=\boldsymbol{\mu}_{11}+\boldsymbol{\varepsilon}_{11}$ and $\boldsymbol{Y}_0(n_0\times 1)=(Y_{10},\ldots,Y_{n_0 0})'=\boldsymbol{\mu}_{00}+\boldsymbol{\varepsilon}_{00}$ be the sub-sample of $\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_n)'$ comprising the observations of $T_i=1$ and $T_i=0$, respectively. Let $\boldsymbol{\varepsilon}_{11}$ ($\boldsymbol{\varepsilon}_{00}$) be the error vector for the treatment (control) sample and $\boldsymbol{\mu}_{11}$ ($\boldsymbol{\mu}_{00}$) be the conditional mean of $\boldsymbol{Y}_1$ ($\boldsymbol{Y}_0$). We denote $\boldsymbol{u}_{t,i}=(u^{(1)}_{i,j^{(1)}_1},u^{(1)}_{i,j^{(1)}_2},\ldots,u^{(1)}_{i,j^{(1)}_{p_1}})'$, for $i=1,2,\ldots,n_1$, and $\boldsymbol{u}_{c,i}=(u^{(0)}_{i,j^{(0)}_1},u^{(0)}_{i,j^{(0)}_2},\ldots,u^{(0)}_{i,j^{(0)}_{p_0}})'$, for $i=1,2,\ldots,n_0$, as the observational covariates in the treatment and control samples, respectively.

For $t=0,1$, we denote $\boldsymbol{\beta}^{(t)}_l=(\beta^{(t)}_{l1},\ldots,\beta^{(t)}_{lm_{n_t}})'(l=1,2,\ldots,p_t)$, and $\boldsymbol{\beta}_t=(\boldsymbol{\beta}^{(t)'}_1,\ldots,\boldsymbol{\beta}^{(t)'}_{p_t})'$. To obtain the estimation of unknown parameters ν_t and $\boldsymbol{\beta}_t$, we consider the following loss function:

$$L_{n,t}(\nu_t,\boldsymbol{\beta}_t)=\sum_{i=1}^{n_t}\Big[Y_{it}-\nu_t-\sum_{l=1}^{p_t}\sum_{k=1}^{m_{n_t}}\beta^{(t)}_{lk}\phi_k(u^{(t)}_{i,j^{(t)}_l})\Big]^2. \tag{4}$$

For $t=0,1$, the estimator of ν_t and $\boldsymbol{\beta}_t$ can be obtained by minimizing $L_{n,t}(\nu_t,\boldsymbol{\beta}_t)$ with the constraints

$$\sum_{i=1}^{n_t}\sum_{k=1}^{m_{n_t}}\beta^{(t)}_{lk}\phi_k(u^{(t)}_{i,j^{(t)}_l})=0,\quad l=1,\ldots,p_t. \tag{5}$$

These centering constraints are sample analogs of the model identifiability restriction $\mathrm{E}(f_{l,t})=0$ for $l=1,\ldots,p_t$, and $t=0,1$. Similar to Huang et al. (2010), we can convert the constrained optimization to an unconstrained optimization by centering

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

the response and the basis functions. Let

$$\bar{\phi}_{lk}^{(t)} = \frac{1}{n_t}\sum_{i=1}^{n_t}\phi_k(u_{i,j_l^{(t)}}^{(t)}) \quad \text{and} \quad \psi_{lk}^{(t)}(u) = \phi_k(u) - \bar{\phi}_{lk}^{(t)}. \tag{6}$$

Define the value of the centered basis functions of the lth element in $\boldsymbol{u}_i^{(t)}$ as

$$\Pi_{il}^{(t)} = (\psi_{l1}^{(t)}(u_{i,j_l^{(t)}}), \psi_{l2}^{(t)}(u_{i,j_l^{(t)}}), \ldots, \psi_{lm_{n_t}}^{(t)}(u_{i,j_l^{(t)}}))'.$$

For $t = 0, 1$, we denote the total design matrix as $\boldsymbol{\Pi}^{(t)} = (\boldsymbol{\Pi}_1^{(t)}, \boldsymbol{\Pi}_2^{(t)}, \ldots, \boldsymbol{\Pi}_n^{(t)})'$, where $\boldsymbol{\Pi}_i^{(t)} = (\Pi_{i1}^{(t)}, \ldots, \Pi_{ip_t}^{(t)})'$ is the $1 \times m_{n_t}p_t$ "design matrix" corresponding to $\boldsymbol{u}_i^{(t)}$. Let $\boldsymbol{\Pi}_{n_1}^{(1)}$ be submatrix of $\boldsymbol{\Pi}^{(1)}$ corresponding to the treatment sample, and $\boldsymbol{\Pi}_{n_0}^{(0)}$ be submatrix of $\boldsymbol{\Pi}^{(0)}$ corresponding to the control sample, respectively.

With this notation, we can write $L_{n,t}(\nu_t, \beta_{jk}^{(t)})$ as

$$L_{n,t}(\nu_t, \beta_{jk}^{(t)}) = \|\tilde{\boldsymbol{Y}}_t - \boldsymbol{\Pi}_{n_t}^{(t)}\boldsymbol{\beta}_t\|_2^2, \tag{7}$$

where $\tilde{\boldsymbol{Y}}_t = (Y_{1t} - \bar{Y}_t, Y_{2t} - \bar{Y}_t, \ldots, Y_{n_t t} - \bar{Y}_t)$, and $\bar{Y}_t = \sum_{i=1}^{n_t} Y_{it}/n_t$. Then minimizing (4) subject to (5) is equivalent to minimizing (7) with respect to $\boldsymbol{\beta}_t$. For any $t \in \{0, 1\}$, the estimator of $\boldsymbol{\beta}_t$ can be obtained by minimizing (7). This estimation method is similar to that in Huang et al. (2010). By minimizing (7), we can obtain the estimator of $\boldsymbol{\beta}_t$ as

$$\hat{\boldsymbol{\beta}}_t = (\boldsymbol{\Pi}_{n_t}^{(t)'}\boldsymbol{\Pi}_{n_t}^{(t)})^{-}\boldsymbol{\Pi}_{n_t}^{(t)'}\tilde{\boldsymbol{Y}}_t,$$

where $t = 0, 1$, and $(\cdot)^{-}$ denotes a generalized inverse. The matrix $\boldsymbol{\Pi}_{n_t}^{(t)'}\boldsymbol{\Pi}_{n_t}^{(t)}$ will be asymptotically nonsingular under conditions given below, then the generalized inverse is just the standard inverse. With the centering, $\hat{\nu}_t = \bar{Y}_t$ for $t \in \{0, 1\}$. Given $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$, the estimator of $\Delta(\boldsymbol{u}_i)$ for individual i is

$$\hat{\Delta}(\boldsymbol{u}_i) = \hat{\nu}_1 + \sum_{l=1}^{p_1}\sum_{k=1}^{m_{n_1}}\hat{\beta}_{lk}^{(1)}\psi_{lk}^{(1)}(u_{i,j_l^{(1)}}) - \hat{\nu}_0 - \sum_{l=1}^{p_0}\sum_{k=1}^{m_{n_0}}\hat{\beta}_{lk}^{(0)}\psi_{lk}^{(0)}(u_{i,j_l^{(0)}}).$$

Remark 1. When fitting $\mu_1(\boldsymbol{u})$ ($\mu_0(\boldsymbol{u})$) with multiple nonparametric models, we fix the same number of spline interior knots, m_{n_1} (m_{n_0}), for all functional components. To facilitate the implementation, one needs to tune the appropriate values of m_{n_1} and m_{n_0}, which are dependent on the smoothness of the nonparametric functions. In this article, we allow the number of B-spline basis functions, m_{n_1} and m_{n_0}, to be different. The Akaike information criterion (AIC), Bayesian information criterion (BIC), and CV criterion can be adopted to select the values of m_{n_1} and m_{n_0}. Instead of choosing m_{n_1} and m_{n_0}, model averaging is another way to address the unknown m_{n_1} and m_{n_0} issue, which is computationally more time consuming, and will be explored in future work.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

3. MODEL AVERAGING AND THE WEIGHT CHOICE CRITERION

In applications, not all covariates are useful in estimating the treatment effect, and analysts are typically unsure about which covariates should be included in the potential outcome model ex-ante. We suppose a list of candidate models with different choices of the available covariates is considered and use a model averaging method to combine them.

3.1. Model Averaging Estimator

Suppose there are K_n candidate models with distinct specifications of covariates $\boldsymbol{u}$ to approximate (1). Here, we allow the number of candidate models, K_n, to depend on the sample size n. Each candidate model consists of two submodels: treatment model and control model. For the kth candidate model, let $\boldsymbol{\varsigma}^{(k,1)} = (j_1^{(k,1)}, j_2^{(k,1)}, \ldots, j_{l_k^{(1)}}^{(k,1)})'$ ($\boldsymbol{\varsigma}^{(k,0)} = (j_1^{(k,0)}, j_2^{(k,0)}, \ldots, j_{l_k^{(0)}}^{(k,0)})'$) denote the index of variables used in $\mu_1(\boldsymbol{u})(\mu_0(\boldsymbol{u}))$, where $l_k^{(1)}$ ($l_k^{(0)}$) is the number of covariates in treatment (control) model. Thus, the corresponding selections of $\boldsymbol{u}$ are $\boldsymbol{u}^{(k,1)} = (u_{j_1^{(k,1)}}, u_{j_2^{(k,1)}}, \ldots, u_{j_{l_k^{(1)}}^{(k,1)}})'$ and $\boldsymbol{u}^{(k,0)} = (u_{j_1^{(k,0)}}, u_{j_2^{(k,0)}}, \ldots, u_{j_{l_k^{(0)}}^{(k,0)}})'$, respectively. Here, we assume the values of $l_k^{(1)}$ and $l_k^{(0)}$ may differ and are bounded for all $k = 1, \ldots, K_n$. The kth candidate additive model is given by

$$
\begin{aligned}
Y &= [\mu_1^{(k)}(\boldsymbol{u}^{(k,1)}) + \varepsilon^{(k,1)}]\mathbb{I}(T=1) + [\mu_0^{(k)}(\boldsymbol{u}^{(k,0)}) + \varepsilon^{(k,0)}]\mathbb{I}(T=0) \\
&= \left[\nu_1^{(k)} + \sum_{v=1}^{l_k^{(1)}} f_{v,1}^{[k]}(u_{j_v^{(k,1)}}) + \varepsilon^{(k,1)} \right] \mathbb{I}(T=1) \\
&\quad + \left[\nu_0^{(k)} + \sum_{v=1}^{l_k^{(0)}} f_{v,0}^{[k]}(u_{j_v^{(k,0)}}) + \varepsilon^{(k,0)} \right] \mathbb{I}(T=0),
\end{aligned}
$$

where $\varepsilon^{(k,t)} = \varepsilon^{(t)} + \mu_t(\boldsymbol{u}) - [\nu_t^{(k)} + \sum_{v=1}^{l_k^{(t)}} f_{v,t}^{[k]}(u_{j_v^{(k,t)}})]$, for $t = 1, 0$. Note that we allow $\mathrm{E}(\varepsilon^{(k,t)}|\boldsymbol{u}) \neq 0$ in the set of candidate models, which arises when the kth candidate model is misspecified for $\mu_t(\boldsymbol{u})$. To ensure identifiability of the function $f_{v,t}^{[k]}$, we assume $\mathrm{E}(f_{v,t}^{[k]}) = 0$, for $t = 1, 0$, $v = 1, 2, \ldots, l_k^{(t)}$, and $k = 1, 2, \ldots, K_n$.

For the kth approximating model, let $\boldsymbol{\Pi}_{ij}^{(k,1)} = (\psi_1^{(1)}(u_{i,j_j^{(k,1)}}), \ldots, \psi_{m_{n_1}^{(k)}}^{(1)}(u_{i,j_j^{(k,1)}}))'$ and $\boldsymbol{\Pi}_{ij}^{(k,0)} = (\psi_1^{(0)}(u_{i,j_j^{(k,0)}}), \ldots, \psi_{m_{n_0}^{(k)}}^{(0)}(u_{i,j_j^{(k,0)}}))'$ be the values of the centered basis functions at the ith observation of the jth covariate in $\boldsymbol{u}^{(k,1)}$ and $\boldsymbol{u}^{(k,0)}$, respectively. Define $\boldsymbol{\Pi}_{1i}^{(k,1)} = (\boldsymbol{\Pi}_{i1}^{(k,1)'}, \ldots, \boldsymbol{\Pi}_{il_k^{(1)}}^{(k,1)'})$, $\boldsymbol{\Pi}_{0i}^{(k,0)} = (\boldsymbol{\Pi}_{i1}^{(k,0)'}, \ldots, \boldsymbol{\Pi}_{il_k^{(0)}}^{(k,0)'})$, $\boldsymbol{\Pi}_1^{(k)} = (\boldsymbol{\Pi}_{11}^{(k,1)'}, \ldots, \boldsymbol{\Pi}_{1n}^{(k,1)'})'$, and $\boldsymbol{\Pi}_0^{(k)} = (\boldsymbol{\Pi}_{01}^{(k,0)'}, \ldots, \boldsymbol{\Pi}_{0n}^{(k,0)'})'$. Let $\boldsymbol{\Pi}_{1n_1}^{(k)}$ and $\boldsymbol{\Pi}_{0n_0}^{(k)}$ be submatrices of $\boldsymbol{\Pi}_1^{(k)}$ and $\boldsymbol{\Pi}_0^{(k)}$ corresponding to the treatment sample

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

and control sample, respectively. Let $\boldsymbol{\beta}_1^{(k)} = (\boldsymbol{\beta}_1^{(k,1)'}, \boldsymbol{\beta}_2^{(k,1)'}, \ldots, \boldsymbol{\beta}_{l_k^{(1)}}^{(k,1)'})'$ and $\boldsymbol{\beta}_0^{(k)} = (\boldsymbol{\beta}_1^{(k,0)'}, \boldsymbol{\beta}_2^{(k,0)'}, \ldots, \boldsymbol{\beta}_{l_k^{(0)}}^{(k,0)'})'$ be the coefficients for the kth candidate model. Then, the estimator of $\boldsymbol{\beta}_t^{(k)}$ is $\hat{\boldsymbol{\beta}}_t^{(k)} = (\boldsymbol{\Pi}_{tn_t}^{(k)'} \boldsymbol{\Pi}_{tn_t}^{(k)})^{-1} \boldsymbol{\Pi}_{tn_t}^{(k)'} \tilde{\boldsymbol{Y}}_t$, for $t = 0, 1$. Hence, the estimator of the treatment effect $\Delta(\boldsymbol{u}_i)$ under the kth model is equal to

$$\hat{\Delta}^{(k)}(\boldsymbol{u}_i) = \hat{\nu}_1^{(k)} + \boldsymbol{\Pi}_{1i}^{(k,1)'} \hat{\boldsymbol{\beta}}_1^{(k)} - \boldsymbol{\Pi}_{0i}^{(k,0)'} \hat{\boldsymbol{\beta}}_0^{(k)} - \hat{\nu}_0^{(k)}, \tag{8}$$

where $\hat{\nu}_t^{(k)} = \bar{Y}_t$ for $t \in \{0, 1\}$.

The treatment effect vector $\boldsymbol{\Delta} = (\Delta(\boldsymbol{u}_1), \Delta(\boldsymbol{u}_2), \ldots, \Delta(\boldsymbol{u}_n))'$ under the kth candidate model can be estimated as

$$\begin{aligned}
\hat{\boldsymbol{\Delta}}^{(k)} &= (\hat{\Delta}^{(k)}(\boldsymbol{u}_1), \hat{\Delta}^{(k)}(\boldsymbol{u}_2), \ldots, \hat{\Delta}^{(k)}(\boldsymbol{u}_n))' \\
&= \hat{\nu}_1^{(k)} \mathbf{1}_n + \boldsymbol{\Pi}_1^{(k)} \hat{\boldsymbol{\beta}}_1^{(k)} - \boldsymbol{\Pi}_0^{(k)} \hat{\boldsymbol{\beta}}_0^{(k)} - \hat{\nu}_0^{(k)} \mathbf{1}_n \\
&= \bar{Y}_1 \mathbf{1}_n + \boldsymbol{\Pi}_1^{(k)} (\boldsymbol{\Pi}_{1n_1}^{(k)'} \boldsymbol{\Pi}_{1n_1}^{(k)})^{-1} \boldsymbol{\Pi}_{1n_1}^{(k)'} \tilde{\boldsymbol{Y}}_1 - \boldsymbol{\Pi}_0^{(k)} (\boldsymbol{\Pi}_{0n_0}^{(k)'} \boldsymbol{\Pi}_{0n_0}^{(k)})^{-1} \boldsymbol{\Pi}_{0n_0}^{(k)'} \tilde{\boldsymbol{Y}}_0 - \bar{Y}_0 \mathbf{1}_n \\
&= \boldsymbol{P}_1^{(k)} \boldsymbol{Y}_1 - \boldsymbol{P}_0^{(k)} \boldsymbol{Y}_0,
\end{aligned} \tag{9}$$

where $\boldsymbol{P}_t^{(k)} = \boldsymbol{\Pi}_t^{(k)} (\boldsymbol{\Pi}_{tn_t}^{(k)'} \boldsymbol{\Pi}_{tn_t}^{(k)})^{-1} \boldsymbol{\Pi}_{tn_t}^{(k)'} + 1/n_t \mathbf{1}_n \mathbf{1}_{n_t}' - 1/n_t \boldsymbol{\Pi}_t^{(k)} (\boldsymbol{\Pi}_{tn_t}^{(k)'} \boldsymbol{\Pi}_{tn_t}^{(k)})^{-1} \boldsymbol{\Pi}_{tn_t}^{(k)'}) \mathbf{1}_n \mathbf{1}_{n_t}'$ for $t = 0, 1$, and $\mathbf{1}_n$ is an $n \times 1$ vector with all elements being 1.

With the treatment effect estimator of each candidate model readily available, the model averaging estimator of $\boldsymbol{\Delta}$ is given by

$$\begin{aligned}
\hat{\boldsymbol{\Delta}}(\boldsymbol{\omega}) &= \sum_{k=1}^{K_n} \omega_k \hat{\boldsymbol{\Delta}}^{(k)} \\
&= \sum_{k=1}^{K_n} \omega_k \boldsymbol{P}_1^{(k)} \boldsymbol{Y}_1 - \sum_{k=1}^{K_n} \omega_k \boldsymbol{P}_0^{(k)} \boldsymbol{Y}_0 \\
&= \boldsymbol{P}_1(\boldsymbol{\omega}) \boldsymbol{Y}_1 - \boldsymbol{P}_0(\boldsymbol{\omega}) \boldsymbol{Y}_0,
\end{aligned} \tag{10}$$

where $\boldsymbol{P}_1(\boldsymbol{\omega}) = \sum_{k=1}^{K_n} \omega_k \boldsymbol{P}_1^{(k)}$, $\boldsymbol{P}_0(\boldsymbol{\omega}) = \sum_{k=1}^{K_n} \omega_k \boldsymbol{P}_0^{(k)}$, and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_{K_n})'$ is the weight vector belonging to the set

$$\mathcal{W} = \left\{ \boldsymbol{\omega} \in [0, 1]^{K_n}, \sum_{k=1}^{K_n} \omega_k = 1 \right\}.$$

Clearly, the weight vector $\boldsymbol{\omega}$ plays an important role in achieving a good and reliable estimation performance. The next section addresses the question of how to determine the weights, $\{\omega_k\}_{k=1}^{K_n}$.

3.2. Weight Choice Criterion

Our weight choice aims at minimizing the squared loss function, $L_n(\boldsymbol{\omega}) = \|\hat{\boldsymbol{\Delta}}(\boldsymbol{\omega}) - \boldsymbol{\Delta}\|^2$. However, it is infeasible as a weight selection criterion due to $\boldsymbol{\Delta}$ being

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

unknown. We develop a J-fold ($J > 1$ is a finite positive integer) CV criterion to choose the weight vector $\boldsymbol{\omega}$.

To implement J-fold CV, we divide the dataset into J blocks, so there are $\tilde{M}_n = \lfloor n/J \rfloor$ observations for the first $J-1$ blocks, where $\lfloor \cdot \rfloor$ denotes the floor function, and $n-(J-1)\tilde{M}_n$ observations for the last (Jth) block. For the kth candidate model, we denote $\tilde{\boldsymbol{\beta}}_t^{(k),[-j]}$ and $\tilde{\nu}_t^{[-j]}$ as the estimators of $\boldsymbol{\beta}_t^{(k)}$ and ν_t without using the observations from the jth block for $j = 1, 2, \ldots, J$, and $t = 0, 1$. Moreover, let $\boldsymbol{\Pi}_{1k}^{[j]}$ and $\boldsymbol{\Pi}_{0k}^{[j]}$ be the covariate matrix corresponding to the subsample of the jth block for the treatment and control model in the kth candidate model, respectively. Then, for the kth candidate model, the corresponding leave-block-out treatment effect estimator is

$$\begin{aligned}
\tilde{\boldsymbol{\Delta}}_k &= \begin{pmatrix} \tilde{\nu}_1^{[-1]}\mathbf{1}_{\tilde{M}_n} \\ \tilde{\nu}_1^{[-2]}\mathbf{1}_{\tilde{M}_n} \\ \vdots \\ \tilde{\nu}_1^{[-J]}\mathbf{1}_{n-(J-1)\tilde{M}_n} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Pi}_{1k}^{[1]} & & & \\ & \boldsymbol{\Pi}_{1k}^{[2]} & & \\ & & \ddots & \\ & & & \boldsymbol{\Pi}_{1k}^{[J]} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}}_1^{(k),[-1]} \\ \tilde{\boldsymbol{\beta}}_1^{(k),[-2]} \\ \vdots \\ \tilde{\boldsymbol{\beta}}_1^{(k),[-J]} \end{pmatrix} \\
&\quad - \begin{pmatrix} \tilde{\nu}_0^{[-1]}\mathbf{1}_{\tilde{M}_n} \\ \tilde{\nu}_0^{[-2]}\mathbf{1}_{\tilde{M}_n} \\ \vdots \\ \tilde{\nu}_0^{[-J]}\mathbf{1}_{n-(J-1)\tilde{M}_n} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\Pi}_{0k}^{[1]} & & & \\ & \boldsymbol{\Pi}_{0k}^{[2]} & & \\ & & \ddots & \\ & & & \boldsymbol{\Pi}_{0k}^{[J]} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{\beta}}_0^{(k),[-1]} \\ \tilde{\boldsymbol{\beta}}_0^{(k),[-2]} \\ \vdots \\ \tilde{\boldsymbol{\beta}}_0^{(k),[-J]} \end{pmatrix} \\
&= \tilde{\boldsymbol{\nu}}_{(1)}^{(k)} - \tilde{\boldsymbol{\nu}}_{(0)}^{(k)} + \boldsymbol{A}_k^{(1)} \begin{pmatrix} \tilde{\boldsymbol{\beta}}_1^{(k),[-1]} \\ \tilde{\boldsymbol{\beta}}_1^{(k),[-2]} \\ \vdots \\ \tilde{\boldsymbol{\beta}}_1^{(k),[-J]} \end{pmatrix} - \boldsymbol{A}_k^{(0)} \begin{pmatrix} \tilde{\boldsymbol{\beta}}_0^{(k),[-1]} \\ \tilde{\boldsymbol{\beta}}_0^{(k),[-2]} \\ \vdots \\ \tilde{\boldsymbol{\beta}}_0^{(k),[-J]} \end{pmatrix}, \qquad (11)
\end{aligned}$$

where $\boldsymbol{A}_k^{(t)} = \text{diag}\{\boldsymbol{\Pi}_{tk}^{[1]}, \boldsymbol{\Pi}_{tk}^{[2]}, \ldots, \boldsymbol{\Pi}_{tk}^{[J]}\}$, and $\tilde{\boldsymbol{\nu}}_{(t)}^{(k)} = (\tilde{\nu}_t^{[-1]}\mathbf{1}'_{\tilde{M}_n}, \tilde{\nu}_t^{[-2]}\mathbf{1}'_{\tilde{M}_n}, \ldots, \tilde{\nu}_t^{[-J]}\mathbf{1}_{n-(J-1)\tilde{M}_n}{}')'$, for $t = 0, 1$. Then, the averaging leave-block-out estimator of $\boldsymbol{\Delta}$ is given by

$$\tilde{\boldsymbol{\Delta}}(\boldsymbol{\omega}) = \sum_{k=1}^{K_n} \omega_k \tilde{\boldsymbol{\Delta}}^{(k)}. \qquad (12)$$

Then, we may attempt to use the J-fold CV criterion to select $\boldsymbol{\omega}$ by

$$\tilde{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega} \in \mathcal{W}} \widetilde{\text{CV}}_n(\boldsymbol{\omega}),$$

where

$$\widetilde{\text{CV}}_n(\boldsymbol{\omega}) = \|\boldsymbol{Y}(1) - \boldsymbol{Y}(0) - \tilde{\boldsymbol{\Delta}}(\boldsymbol{\omega})\|^2.$$

However, a key challenge is that, different from a standard regression problem, it is not implementable because we do not observe $\boldsymbol{Y}(1) - \boldsymbol{Y}(0)$. One possible

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

solution, as suggested in Rolling and Yang (2014) and Rolling et al. (2019), is to adopt a matching scheme to estimate the unobserved potential outcome for each individual. The matching estimator calculates the unobserved potential outcomes as

$$\hat{Y}_i(0) = \begin{cases} Y_i, & \text{if} \quad T_i = 0, \\ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j, & \text{if} \quad T_i = 1, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j, & \text{if} \quad T_i = 0, \\ Y_i, & \text{if} \quad T_i = 1, \end{cases}$$

where $\mathcal{J}_M(i)$ denotes the set of indices for the first M matches for unit i, which is $\mathcal{J}_M(i) = \{j_1(i), \ldots, j_M(i)\}$. Here, $j_m(i)$ is the index of the individual who is the mth closest to unit i among the individuals with the treatment value opposite to that of individual i. The Mahalanobis distance of the covariate between two individuals is applied to measure the proximity. When we use a matching estimator to substitute for the unobserved potential outcome, the difference between potential outcomes $Y_i(1) - Y_i(0)$ can be approximated by $\hat{Y}_i(1) - \hat{Y}_i(0)$. Therefore, we have

$$\begin{aligned}
&\hat{Y}_i(1) - \hat{Y}_i(0) \\
&= \left\{ Y_i - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j \right\} T_i + \left\{ \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} Y_j - Y_i \right\} (1 - T_i) \\
&= \mu_1(\boldsymbol{u}_i) - \mu_0(\boldsymbol{u}_i) + (2T_i - 1)\frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} (\mu_{1-T_i}(\boldsymbol{u}_i) - \mu_{1-T_i}(\boldsymbol{u}_j)) \\
&\quad + (2T_i - 1)\left(\varepsilon_i^{(T_i)} - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} \varepsilon_j^{(1-T_i)} \right) \\
&= \Delta(\boldsymbol{u}_i) + b_{i,M} + \eta_{i,M}, \qquad \textbf{(13)}
\end{aligned}$$

where

$$b_{i,M} = \frac{1}{M}(2T_i - 1)\sum_{j\in\mathcal{J}_M(i)} (\mu_{1-T_i}(\boldsymbol{u}_i) - \mu_{1-T_i}(\boldsymbol{u}_j))$$

is the conditional bias arising from the approximation of $\Delta(\boldsymbol{u}_i)$ by $\hat{Y}_i(1) - \hat{Y}_i(0)$, and

$$\eta_{i,M} = (2T_i - 1)\left(\varepsilon_i^{(T_i)} - \frac{1}{M}\sum_{j\in\mathcal{J}_M(i)} \varepsilon_j^{(1-T_i)} \right)$$

is the residual item.

Let $(\hat{Y}_1(1) - \hat{Y}_1(0), \ldots, \hat{Y}_n(1) - \hat{Y}_n(0))' = \boldsymbol{\Delta} + \boldsymbol{b}_M + \boldsymbol{\eta}_M$ be the matching estimator, where $\boldsymbol{b}_M = (b_{1,M}, \ldots, b_{n,M})'$ is the bias vector and $\boldsymbol{\eta}_M = (\eta_{1,M}, \ldots, \eta_{n,M})'$

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

is the residual vector. We assume that the value of M in the matching estimator is a bounded positive integer that is independent of the sample size n. When $M = 1$, it is the nearest neighbor matching estimator. As mentioned by Abadie and Imbens (2011), matching estimators are often implemented with small values for M, as small as 1. Therefore, in accordance with the methodologies employed by Rolling and Yang (2014), Rolling et al. (2019), and Zhao et al. (2024), we utilize the nearest neighbor matching ($M = 1$) for subsequent simulation studies and empirical applications.

It is worth noting that Rolling and Yang (2014) and Zhao et al. (2024) apply a partition method to construct potential outcomes because independence is needed in their theoretical analysis. However, they adopt nearest neighbor matching to replace the partition method in their simulations for computational convenience. Unlike Rolling and Yang (2014) and Zhao et al. (2024), our theoretical and simulation studies are all based on the same matching method. In particular, we study the theoretical properties of bias item $\boldsymbol{b}_M$ and residual item $\boldsymbol{\eta}_M$. This is one of the technical contributions of our work.

When the unobserved potential outcomes are imputed by the matching estimates, the J-fold CV criterion for choosing the weight vector $\boldsymbol{\omega}$ is defined as

$$\mathrm{CV}_n(\boldsymbol{\omega}) = \|\hat{\boldsymbol{\tau}} - \tilde{\boldsymbol{\Delta}}(\boldsymbol{\omega})\|^2, \tag{14}$$

where $\hat{\boldsymbol{\tau}} = (\hat{Y}_1(1) - \hat{Y}_1(0), \ldots, \hat{Y}_n(1) - \hat{Y}_n(0))'$. The optimal weight vector is obtained by choosing $\boldsymbol{\omega} \in \mathcal{W}$ to minimize $\mathrm{CV}_n(\boldsymbol{\omega})$, i.e.,

$$\hat{\boldsymbol{\omega}} = \arg\min_{\boldsymbol{\omega}\in\mathcal{W}} \mathrm{CV}_n(\boldsymbol{\omega}).$$

Then, the model averaging estimator of $\boldsymbol{\Delta}$ is $\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\omega}})$. In the simulation studies, we find that the choice of J generally has little impact on the final results, so we choose $J = 5$ or 10.

Remark 2. When the dimension of covariates is particularly large, the number of all possible candidate models is huge, which not only increases the computational burden but also may make the combination strategy infeasible. In this case, we can implement a model-screening step before averaging. Since the choice of the candidate models can affect the final performance, we should prepare a candidate model set carefully. In the spirit of "sure independence screening" proposed by Fan, Feng, and Song (2011), we order the covariates based on their marginal correlations with the response variable in treatment (control) sample and construct the set of the treatment (control) candidate models by including one additional covariate each time based on the ordering. The final candidate model set comprises all possible combinations of the treatment candidate models and the control candidate models. Similar screening procedures have been used in other model-averaging studies, such as Claeskens, Croux, and Van Kerckhoven (2006) and Zhang et al. (2016). This strategy is taken in our simulation, and it works very well as will be seen later.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Remark 3. In most econometric applications, some variables in $\boldsymbol{U}$ may be discrete. Terms involving categorical variables in additive models can be set as linear (the ordinal case) or indicator functions (the nominal case). Thus, our proposed method can handle mixed-type datasets comprising both continuous and categorical predictors.

4. ASYMPTOTIC PROPERTIES

In this section, we establish the theoretical properties of $\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\omega}})$ and $\hat{\boldsymbol{\omega}}$. We first show that the proposed estimator is asymptotically optimal when the candidate models are all misspecified. Here, "optimal" means that its estimation error is asymptotically identical to that of the infeasible but best model averaging estimator. Also, we study the consistency properties of $\hat{\boldsymbol{\omega}}$ and $\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\omega}})$ when the set of candidate models includes correct models. We say the s-th model is a correct model in terms of including all the true variables in the regression functions if and only if there exist two vectors $\boldsymbol{f}_1^{[s]} = (\nu_1^{(s)}, f_{1,1}^{[s]}, f_{2,1}^{[s]}, \ldots, f_{l_s^{(1)},1}^{[s]})'$ and $\boldsymbol{f}_0^{[s]} = (\nu_0^{(s)}, f_{1,0}^{[s]}, f_{2,0}^{[s]}, \ldots, f_{l_s^{(0)},0}^{[s]})'$, such that $\mu_1(\boldsymbol{u}) = \nu_1^{(s)} + \sum_{v=1}^{l_s^{(1)}} f_{v,1}^{[s]}(u_{j_v^{(s,1)}})$ and $\mu_0(\boldsymbol{u}) = \nu_0^{(s)} + \sum_{v=1}^{l_s^{(0)}} f_{v,0}^{[s]}(u_{j_v^{(s,0)}})$. A correct model is also called a *correctly specified* model, and the opposite is a *misspecified* model. Without loss of generality, we assume that the first $S_0 \geq 1$ models are correct when correct models are available in the candidate set. In this context, we postulate that the number of correct models remains bounded, notwithstanding the potential divergence in the number of candidate models.

4.1. Asymptotic Optimality

In this section, we establish the asymptotic optimality of the proposed model averaging estimator when all candidate models are misspecified. Define the squared error as

$$L_n(\boldsymbol{\omega}) = \|\hat{\boldsymbol{\Delta}}(\boldsymbol{\omega}) - \boldsymbol{\Delta}\|^2. \tag{15}$$

For the kth candidate model, $\hat{\boldsymbol{\beta}}_1^{(k)}$ and $\hat{\boldsymbol{\beta}}_0^{(k)}$ are the solutions to the following score equations:

$$\boldsymbol{\Pi}_{1n_1}^{(k)'}(\tilde{\boldsymbol{Y}}_1 - \boldsymbol{\Pi}_{1n_1}^{(k)}\boldsymbol{\beta}_1^{(k)}) = 0 \quad \text{and} \quad \boldsymbol{\Pi}_{0n_0}^{(k)'}(\tilde{\boldsymbol{Y}}_0 - \boldsymbol{\Pi}_{0n_0}^{(k)}\boldsymbol{\beta}_0^{(k)}) = 0.$$

By the definition of "quasi-true" parameters in Lv and Liu (2014), we assume that $\boldsymbol{\beta}_1^{(k)*} = (\boldsymbol{\beta}_{11}^{(k)*'}, \boldsymbol{\beta}_{21}^{(k)*'}, \ldots, \boldsymbol{\beta}_{l_k^{(1)}1}^{(k)*'})'$ and $\boldsymbol{\beta}_0^{(k)*} = (\boldsymbol{\beta}_{10}^{(k)*'}, \boldsymbol{\beta}_{20}^{(k)*'}, \ldots, \boldsymbol{\beta}_{l_k^{(0)}0}^{(k)*'})'$ are, respectively, the solutions to the equations:

$$\boldsymbol{\Pi}_{1n_1}^{(k)'}\left[\left(\boldsymbol{I}_{n_1} - \frac{1}{n_1}\mathbf{1}_{n_1}\mathbf{1}_{n_1}'\right)\boldsymbol{\mu}_{11} - \boldsymbol{\Pi}_{1n_1}^{(k)}\boldsymbol{\beta}_1^{(k)}\right] = 0$$

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

and

$$\mathbf{\Pi}_{0n_0}^{(k)'}\left[\left(\boldsymbol{I}_{n_0}-\frac{1}{n_0}\mathbf{1}_{n_0}\mathbf{1}'_{n_0}\right)\boldsymbol{\mu}_{00}-\mathbf{\Pi}_{0n_0}^{(k)}\boldsymbol{\beta}_0^{(k)}\right]=0,$$

where $\boldsymbol{I}_n$ denotes an $n\times n$ identity matrix. The convergence order of $\hat{\boldsymbol{\beta}}_1^{(k)}$ and $\hat{\boldsymbol{\beta}}_0^{(k)}$ between "quasi-true" parameters $\boldsymbol{\beta}_1^{(k)*}$ and $\boldsymbol{\beta}_0^{(k)*}$ in the l_2-norm will be discussed in a lemma to be given soon. In addition, by the law of large numbers, we can show that ν_1 and ν_0 are the limit values of $\hat{\nu}_1^{(k)}$ and $\hat{\nu}_0^{(k)}$, respectively. Hence, for given $\boldsymbol{\beta}_1^{(k)*}$ and $\boldsymbol{\beta}_0^{(k)*}$, we know $\mathbf{\Delta}^{(k)*}=\nu_1\mathbf{1}_n+\mathbf{\Pi}_1^{(k)}\boldsymbol{\beta}_1^{(k)*}-\mathbf{\Pi}_0^{(k)}\boldsymbol{\beta}_0^{(k)*}-\nu_0\mathbf{1}_n$. Hence, the corresponding loss function is defined as

$$L_n^*(\boldsymbol{\omega})=\|\mathbf{\Delta}^*(\boldsymbol{\omega})-\mathbf{\Delta}\|^2,$$

where $\mathbf{\Delta}^*(\boldsymbol{\omega})=\sum_{k=1}^{K_n}\omega_k\mathbf{\Delta}^{(k)*}$.

Let $\xi_n=\inf_{\boldsymbol{\omega}\in\mathcal{W}}L_n^*(\boldsymbol{\omega})$, $m_n=n^{1/(2d+1)}$, $m_{n_t}^{\star}=\max_{1\le k\le K_n}m_{n_t}^{(k)}$, and $\check{m}_{n_t}=\min_{1\le k\le K_n}m_{n_t}^{(k)}$ for $t=0,1$. Let $\boldsymbol{\omega}_k^{(0)}$ be a $K_n\times 1$ vector in which the kth element is one and all others are zero. The following regularity conditions are required for our results. All limiting statements, unless stated otherwise, correspond to $n\to\infty$.

Condition 3. Let $\boldsymbol{U}$ be a $p\times 1$ of continuous covariates with density function bounded above and away from zero on $\mathbb{U}$.

Condition 4. (i) $\mu_1(\boldsymbol{u})$ and $\mu_0(\boldsymbol{u})$ are Lipschitz and uniformly bounded in $\mathbb{U}$, (ii) $\sigma_1^2(\boldsymbol{u})$ and $\sigma_0^2(\boldsymbol{u})$ are Lipschitz and uniform bounded above and away from zero, and (iii) $\mathrm{E}(Y^4|\boldsymbol{U}=\boldsymbol{u},T=1)$ and $\mathrm{E}(Y^4|\boldsymbol{U}=\boldsymbol{u},T=0)$ are uniformly bounded in $\boldsymbol{u}\in\mathbb{U}$ almost surely.

Condition 5. Let κ be a nonnegative integer, and let $\alpha\in(0,1]$ be such that $d=\kappa+\alpha>0.5$. For $t=0,1$ and $1\le j\le p_t$, we assume the function $f_{j,t}$ on $[a,b]$ has the κth derivative $f_{j,t}^{(\kappa)}$ that satisfies a Lipschitz condition of the order α:

$$\left|f_{j,t}^{(\kappa)}(s_1)-f_{j,t}^{(\kappa)}(s_2)\right|\le C|s_1-s_2|^{\alpha}\qquad\text{for}\quad s_1,s_2\in[a,b],$$

where C is a positive constant.

Condition 6. $\check{m}_{n_t}\asymp n_t^{1/(2d+1)}$ and $m_{n_t}^{\star}\asymp n_t^{1/(2d+1)}$, for $t=0,1$.

Condition 7. There exists a positive constant C_0 such that

$$\max_{t\in\{0,1\}}\max_{1\le k\le K_n}\frac{\|\boldsymbol{\beta}_t^{(k)*}\|^2}{m_{n_t}^{(k)}}\le C_0. \tag{16}$$

Condition 8. (i) $\xi_n^{-1}n^{1-2/p}=o(1)$ and (ii) $\xi_n^{-1}n^{1/2}K_n^2=o(1)$.

Conditions 3–5 are standard conditions for nonparametric additive models. Condition 3 is the restriction imposed on the covariates, as is commonly-used in additive modeling, such as Condition (A4) in Huang et al. (2010). Condition 4 requires that $\mu_t(\boldsymbol{u})$ and $\sigma_t^2(\boldsymbol{u})$, for $t=0,1$, have a strong continuous property, and

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

it implies that $\{\Delta(\boldsymbol{u}_i)\}_{i=1}^n$ are bounded by some constant. It is used in matching estimator research (Abadie and Imbens, 2006; Abadie and Imbens, 2011) to study asymptotic properties of the bias vector $\boldsymbol{b}_M$ and residual vector $\boldsymbol{\eta}_M$ in the matching estimator. Condition 6 is a sensible assumption about the number of B-spline basis functions. Condition 7 restricts the variability of the coefficients. It resembles Condition 3 of Zou et al. (2022). Condition 5 is a typical smoothness condition on the nonparametric functions $f_{j,t}(\cdot)$, for $t = 0, 1$ and $1 \le j \le p_t$.

Condition 8 requires that ξ_n grows at a rate no slower than $n^{1/2}K_n^2$ and $n^{1-2/p}$. Note that $\xi_n \to \infty$ is obviously necessary for Condition 8 to hold, which indicates that there is no candidate model for which the bias is zero. To better understand Condition 8, we consider a situation where the set of candidate models includes at least one correct model. If we denote s as the index of a correct model, then we have $\mu_1(\boldsymbol{u}) = \nu_1^{(s)} + \sum_{v=1}^{l_s^{(1)}} f_{v,1}^{[s]}(u_{j_v^{(s,1)}})$ and $\mu_0(\boldsymbol{u}) = \nu_0^{(s)} + \sum_{v=1}^{l_s^{(0)}} f_{v,0}^{[s]}(u_{j_v^{(s,0)}})$. Note that in this case

$$\xi_n = \inf_{\boldsymbol{\omega}\in\mathcal{W}} L_n^*(\boldsymbol{\omega}) \le \|\boldsymbol{\Delta}^{(s)*} - \boldsymbol{\Delta}\|^2 = O_p(nm_n^{-2d}),$$

where the last equality can be deduced from Lemma A.3 in the Supplementary Material. From $d = \kappa + \alpha > 0.5$, we observe that $n^{1/2}K_n^2/(nm_n^{-2d})$ approaches infinity, not 0, implying that Condition 8 is violated. Hence, Condition 8 requires all candidate models to be misspecified. Moreover, we observe that $\|\boldsymbol{b}_M\|^2 = O_p(n^{1-2/p})$ in Lemma 1. Therefore, the first part of Condition 8 assumes that the growth rate of $\|\boldsymbol{b}_M\|^2$ must be slower than the rate of ξ_n. This is a common condition in most treatment effect model averaging literature (Zhao et al., 2024; Cui et al., 2025), and so is $\xi_n^{-2}nK_n^2m_n = o(1)$ (see, e.g., Zhang, Zou, and Liang, 2014, Condition (A7); Zhang et al., 2016, Condition (C.3)).

To study the asymptotic optimality of our proposed procedure, the behaviors of the bias vector $\boldsymbol{b}_M$ and residual vector $\boldsymbol{\eta}_M$ need to be understood.

LEMMA 1. *Under Conditions 1–4, we have*

$$\|\boldsymbol{b}_M\| = O_p(n^{1/2-1/p}), \quad \textit{and} \quad \|\boldsymbol{\eta}_M\| = O_p(n^{1/2}). \tag{17}$$

LEMMA 2. *Suppose that Conditions 2–6 hold. We have*

$$\max_{1\le k\le K_n} \|\hat{\boldsymbol{\beta}}_1^{(k)} - \boldsymbol{\beta}_1^{(k)*}\|^2 = O_p\left(\frac{K_n m_n}{n}\right) \tag{18}$$

and

$$\max_{1\le k\le K_n} \|\hat{\boldsymbol{\beta}}_0^{(k)} - \boldsymbol{\beta}_0^{(k)*}\|^2 = O_p\left(\frac{K_n m_n}{n}\right). \tag{19}$$

Lemma 2 states that the estimators $\hat{\boldsymbol{\beta}}_1^{(k)}$ and $\hat{\boldsymbol{\beta}}_0^{(k)}$ converge to the "quasi-true" parameters, $\boldsymbol{\beta}_1^{(k)*}$ and $\boldsymbol{\beta}_0^{(k)*}$, of the kth candidate model at the uniform speed of $K_n m_n n^{-1}$. It is vital for establishing the following asymptotic optimality.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

THEOREM 1. *Under the conditions of Lemmas 1 and 2, and Conditions 7 and 8, we have*

$$\frac{L_n(\hat{\boldsymbol{\omega}})}{\inf_{\boldsymbol{\omega}\in\mathcal{W}} L_n(\boldsymbol{\omega})} \to 1 \tag{20}$$

in probability as $n \to \infty$.

Theorem 1 presents the asymptotic optimality of the weight selection criterion. It shows a theoretical advantage of our method over other model selection and model averaging methods since the loss of the proposed method is asymptotically identical to that of the best possible model averaging estimator. The proof of Theorem 1 and the necessary lemmas are given in the Supplementary Material.

4.2. Consistency of Weights and the Model Averaging Estimator

In this section, we consider a different situation where some correct models are available in the candidate model set. We show that our proposed procedure asymptotically assigns all weight to the correctly specified models. Moreover, we show that the proposed CATE estimator converges to the true CATE when the set of candidate models includes at least one correct model.

Let $\hat{\Gamma} = \sum_{k=1}^{S_0} \hat{\omega}_k$ be the sum of weights put on the correctly specified models, where $\hat{\omega}_k$ is the kth element of the weight vector $\boldsymbol{\omega}$. Let $\mathcal{W}_F = \{\omega \in \mathcal{W} : \omega_k = 0 \text{ for } k = 1, 2, \ldots, S_0\}$ denote the subset of $\mathcal{W}$ that assigns zero weight to the correctly specified models. Let $\xi_{n,F} = \inf_{\boldsymbol{\omega}\in\mathcal{W}_F} L_n^*(\boldsymbol{\omega})$ be the infimum of $L_n^*(\boldsymbol{\omega})$ on the set $\mathcal{W}_F$. Let $\tilde{\boldsymbol{\omega}}_F = (\hat{\omega}_{S_0+1}, \ldots, \hat{\omega}_{K_n})'$ and $\Delta_j^{(k)*}$ be the jth component of $\boldsymbol{\Delta}^{(k)*}$. Define $\boldsymbol{\zeta}_i = (\Delta_i^{(S_0+1)*} - \Delta_i^{(S_0)*}, \Delta_i^{(S_0+2)*} - \Delta_i^{(S_0)*}, \ldots, \Delta_i^{(K_n)*} - \Delta_i^{(S_0)*})'$, and

$$\boldsymbol{\Psi}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\zeta}_i \boldsymbol{\zeta}_i'.$$

We aim to show the consistency of the averaging weights and model averaging estimation under the following extra conditions.

Condition 9.

$$\xi_{n,F}^{-1/2} n^{1/2-1/p} = o(1) \quad \text{and} \quad \xi_{n,F}^{-1} n^{1/2} K_n^2 = o(1).$$

Condition 10. For sufficiently large n, $\underline{\lambda}(\boldsymbol{\Psi}_n) \geq C_2$ for some positive constants C_2.

Condition 11. $(K_n - S_0)(m_n^{1/2} \vee K_n^{1/2} n^{1/2-1/p}) n^{-1/2} = o(1)$, and $K_n^3/m_n = o(1)$.

Condition 9 is adopted to naturally distinguish between the misspecified and correct models. The infimum of the squared loss $\xi_{n,F}$, which is based on the average estimator of the misspecified models, is required to have a sufficiently large divergence rate. In particular, $\xi_{n,F}$ grows no slower than $n^{1-2/p}$ and $n^{1/2}K_n^2$.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Condition 10 requires some restrictions on the correct models and on misspecified models as well. More specifically, it demands that the difference between the misspecified models and the S_0-th correct model cannot be negligible asymptotically (Racine et al., 2023). Condition 11 restricts on the relative divergence rates of m_n, n, and K_n. If $K_n = O(1)$, then by $m_n = n^{1/(2d+1)}$, we observe $(K_n - S_0)m_n^{1/2}n^{-1/2} = n^{-d/(2d+1)} = o(1)$ and $(K_n - S_0)m_n^{1/2}n^{-1/2} = n^{-1/p} = o(1)$. If we assume K_n diverges to ∞ at the rate of $K_n = O(n^{\frac{d-\delta_1}{2d+1}} \wedge n^{(\frac{2}{3p}-\delta_2)})$ with $0 < \delta_1 < d$, $0 < \delta_2 < \frac{2}{3p}$, and $S_0 = O(1)$, then by $m_n = n^{1/(2d+1)}$, we also have

$$(K_n - S_0)m_n^{1/2}n^{-1/2} = n^{-\delta_1/(2d+1)} = o(1)$$

and

$$(K_n - S_0)K_n^{1/2}n^{1/2-1/p}n^{-1/2} = o(1).$$

Thus, Condition 11 is plausible.

LEMMA 3. *Suppose Conditions 2–6 are satisfied. For any $k = 1, 2, \ldots, S_0$, we then have*

$$\max_{1\le i\le n} \|\hat{\Delta}^{(k)}(\boldsymbol{u}_j) - \Delta(\boldsymbol{u}_j)\|^2 = O_p(m_n^{-2d}) = O_p(n^{-2d/(2d+1)}). \tag{21}$$

Equation (21) gives the consistency rate of the estimator for the nonparametric additive model when the kth candidate model is correct, similar to Corollary 1 in Stone (1985) and Corollary 2 in Huang et al. (2010). Moreover, for any $k = 1, 2 \ldots, S_0$, equation (21) implies that $\|\hat{\boldsymbol{\Delta}}^{(k)} - \boldsymbol{\Delta}\|^2 = O_p(nm_n^{-2d})$.

THEOREM 2. *Under the conditions of Lemmas 1–3, and Conditions 7, 8, and 10, we have that*

$$\hat{\Gamma} \to 1 \tag{22}$$

in probability as $n \to \infty$.

Theorem 2 shows that our method tends to assign all weights to the correct models when they are included. This result, combined with the asymptotic optimality, demonstrates the effectiveness of our method. Next, we give the convergence rate of the weights of the misspecified models.

LEMMA 4. *If the conditions of Lemmas 1–3, and Conditions 7, 10, and 11 are satisfied, then*

$$\|\tilde{\boldsymbol{\omega}}_F\| = O_p\left(\frac{\sqrt{K_n - S_0}m_n^{1/2}}{n^{1/2}} + \frac{\sqrt{K_n - S_0}K_n^{1/2}n^{1/2-1/p}}{n^{1/2}}\right).$$

Lemma 4 shows that the weights assigned to the misspecified models converge to zero at a certain rate. If the number of candidate models is limited and $m_n =$

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

$n^{1/(2d+1)}$, we have $\|\tilde{\boldsymbol{\omega}}_F\| = O_p(m_n^{1/2}/n^{1/2} + n^{-1/p}) = O_p(n^{-d/(2d+1)-1/p})$. If we assume K_n diverges to ∞ at the rate of $n^{1/p-\delta_1}$ with $0 < \delta_1 < 1/p$, and $m_n^{1/2} = O_p(K_n^{1/2} n^{1/2-1/p})$, we have

$$\|\tilde{\boldsymbol{\omega}}_F\| = O_p\left(\frac{\sqrt{K_n - S_0} K_n^{1/2} n^{1/2-1/p}}{n^{1/2}}\right) = O_p\left(n^{-\delta_1}\right).$$

This property plays a pivotal role in the derivation of the consistency of $\hat{\Delta}_i(\hat{\boldsymbol{\omega}})$, where $\hat{\Delta}_i(\hat{\boldsymbol{\omega}})$ is the ith component of $\hat{\boldsymbol{\Delta}}(\hat{\boldsymbol{\omega}})$.

We now state a uniform consistency of our proposed model averaging estimator.

THEOREM 3. *Suppose that the conditions in Lemma 4 are satisfied. If there exist two positive constants κ_1 and κ_2 such that $E(\|\boldsymbol{Y}_1\|_\infty) \le \kappa_1 < \infty$ and $E(\|\boldsymbol{Y}_0\|_\infty) \le \kappa_1 < \infty$, then we have*

$$\max_{1\le i\le n} \|\hat{\Delta}_i(\hat{\boldsymbol{\omega}}) - \Delta(\boldsymbol{u}_i)\|^2 = O_p\left(\frac{(K_n - S_0)^2 m_n}{n} + \frac{(K_n - S_0)^2 K_n n^{1-2/p}}{n}\right).$$

Remark 4. Theorem 3 shows that our proposed model averaging estimator will tend to the true CATE under some regularity conditions. Typically, (i) if $K_n = O(1)$, by $m_n = n^{1/(2d+1)}$, then we have

$$\max_{1\le i\le n} \|\hat{\Delta}_i(\hat{\boldsymbol{\omega}}) - \Delta(\boldsymbol{u}_i)\|^2 = O_p(m_n^{-2d} + n^{-2/p}).$$

(ii) If $K_n = O(n^{\frac{2}{3p}-\delta_1})$ with $0 < \delta_1 < 2/(3p)$, and $m_n = O_p(K_n n^{1-2/p})$, then we observe

$$\max_{1\le i\le n} \|\hat{\Delta}_i(\hat{\boldsymbol{\omega}}) - \Delta(\boldsymbol{u}_i)\|^2 = O_p(n^{-\delta_1}).$$

4.3. Sufficient Dimension Reduction for Matching

When p is moderate or large, for a given observation i, it may be difficult to identify an observation in the other treatment group that possesses a covariate value similar to $\boldsymbol{U}_i$. Thus, our method may suffer from the so-called "curse of dimensionality," which makes matching ineffective. To address this issue, nonlinear sufficient dimension reduction is applied to reduce the dimension of the covariates prior to matching. The idea of nonlinear sufficient dimension reduction is to find low-dimensional nonlinear functions of covariates that hopefully capture all of the information in $\mu_1(\boldsymbol{U}_i)$ and $\mu_0(\boldsymbol{U}_i)$. The nearest neighbor matching based on these low-dimensional nonlinear functions of $\boldsymbol{U}$ becomes much easier to implement and can perform better.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

The nonlinear sufficient dimension reduction is to seek an arbitrary function $\boldsymbol{\varphi}_t = (\varphi_{t,1}, \ldots, \varphi_{t,r(t)}) : \mathbb{R}^p \to \mathbb{R}^{r(t)}$ such that

$$Y(t) \perp \boldsymbol{U} | \boldsymbol{\varphi}_t(\boldsymbol{U}), \quad t \in \{0, 1\},$$

where $\varphi_{t,i} \in L_2(P_{\boldsymbol{U}})$ for any $i = 1, 2, \ldots r(t)$, and $L_2(P_{\boldsymbol{U}})$ is the space of square-integrable functions of $\boldsymbol{U}$. We denote central σ-field as $\mathcal{G}_{Y(t)|\boldsymbol{U}} = \mathcal{F}(\boldsymbol{\varphi}_t)$ ($\mathcal{F}(\boldsymbol{\varphi}_t)$ is the σ-field generated by $\boldsymbol{\varphi}_t$). We use $\mathcal{H}_U(\mathcal{G}_{Y(t)|\boldsymbol{U}})$ to denote the class of all functions g in $L_2(P_{\boldsymbol{U}})$ such that $g(\boldsymbol{U})$ is measurable $\mathcal{G}_{Y(t)|\boldsymbol{U}}$. Under some mild conditions, the sufficient dimension reduction classes $\mathcal{H}_U(\mathcal{G}_{Y(1)|\boldsymbol{U}})$ and $\mathcal{H}_U(\mathcal{G}_{Y(0)|\boldsymbol{U}})$ exist and are unique (Lee, Li, and Chiaromonte, 2013; Luo and Li, 2016; Li and Song, 2017). We denote $\boldsymbol{\varphi}(\boldsymbol{U}_i) = (\boldsymbol{\varphi}_1(\boldsymbol{U}_i), \boldsymbol{\varphi}_0(\boldsymbol{U}_i))'$ for $i = 1, 2, \ldots n$. Let $D(\cdot)$ denote Mahalanobis distance. If $r(1) + r(0) \ll p$, for each subject i in the treatment group, finding a treatment–control pair (i, i^*) such that $D(\boldsymbol{\varphi}(\boldsymbol{U}_i), \boldsymbol{\varphi}(\boldsymbol{U}_{i^*})$ is small is much easier than finding a pair for which $D(\boldsymbol{U}_i, \boldsymbol{U}_{i^*})$ is small. Therefore, when p is moderate or large, the nearest neighbor matching is executed by $\boldsymbol{\varphi}(\boldsymbol{U})$ to impute the unobserved potential outcome for each individual. More specifically, for each subject i in the treatment group, we find M individuals in the control group that correspond to the M smallest value in $\{D(\boldsymbol{\varphi}(\boldsymbol{U}_i), \boldsymbol{\varphi}(\boldsymbol{U}_{i^*})), T_{i^*} = 0\}$. For $t = 0, 1$, because $\mathcal{H}_U(\mathcal{G}_{Y(t)|\boldsymbol{U}})$ is sufficient for $\mu_t(\boldsymbol{U}_i)$, Theorems 1–3 hold if the nearest neighbor matching is implemented by $\boldsymbol{\varphi}(\boldsymbol{U})$, and p is replaced by $q_{\mathcal{S}} = r(1) + r(0)$ (Rolling and Yang, 2014; Luo and Zhu, 2020). Even when the covariate dimension p increases with the sample size, our proposed method and the corresponding theoretical guarantees remain valid provided that $q_{\mathcal{S}}$ is fixed.

In practice, $\mathcal{H}_U(\mathcal{G}_{Y(1)|\boldsymbol{U}})$ and $\mathcal{H}_X(\mathcal{G}_{Y(0)|\boldsymbol{U}})$ are typically unknown, and must be estimated. Li and Song (2017) show that the sufficient dimension reduction class can be consistently estimated by the observed data. There are several methods available to estimate vectors in the central sufficient dimension reduction class without assuming a model for the data, including generalized sliced inverse regression (Lee et al., 2013; Luo and Li, 2016; Li and Song, 2017), principal support vector machine (Li, Artemiou, and Li, 2011), and generalized average variance estimator (Lee et al., 2013; Luo and Li, 2016; Li and Song, 2017). Given the resulting estimates, $\hat{\boldsymbol{\varphi}}(\boldsymbol{U})$, we can use $\hat{\boldsymbol{\varphi}}(\boldsymbol{U})$ in the nearest neighbor matching method mentioned above. In the following simulation studies, we apply generalized sliced inverse regression to estimate vectors in the central sufficient dimension reduction class. Moreover, a CV scheme (Li and Song, 2017) is used to select parameters in generalized sliced inverse regression, and cross-validated BIC (Li and Song, 2017) is applied to determine the values $r(1)$ and $r(0)$.

5. SIMULATION STUDY

In this section, we examine the finite-sample performance of our proposed model averaging method through simulation studies. We compare it with several other methods, including AIC, BIC, treatment effect CV (TECV), smooth-AIC (SAIC),

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Table 1. Model framework for Design 1

	$t=1$	$t=0$
$f_{1,t}$	$5\sin(2\pi u_1)$	$1.5\sin(2\pi u_1)-2u_1$
$f_{2,t}$	$(2u_2+1)^2$	$u_2^3+1.5(u_2+1)^2$
$f_{3,t}$	$5\phi(u_3,0.5,0.8^2)$	$6(0.1\sin(2\pi u_3)+0.2\cos(2\pi u_3)+0.3\sin(4\pi^2u_3^2)$ $+0.4\cos(4\pi^2u_3^2)+0.5\sin(8\pi^3u_3^3))$
$f_{4,t}$	$\exp(2u_4)$	$-\sin(\exp(-0.5u_4))-3u_4^2$
$f_{5,t}$	$3u_5+1.5\exp(\sin(2\pi u_5)+1)$	$-3\exp(u_5^2-u_5)-2\cos(2\pi u_5)$

Note: $\phi(\cdot,0.5,0.8^2)$ is the Gaussian probability density function with a mean of 0.5 and a standard deviation of 0.8.

smooth-BIC (SBIC), TEEM (Rolling et al., 2019), double robust estimator, and causal forest estimator (Wager and Athey, 2018).

5.1. Experimental Designs

To support the theory in Section 4, this section considers four simulation designs with several settings. In the first three designs, all the candidate models are misspecified. In the last, at least one correct model is in the candidate model set.

Design 1: Consider the following data-generating procedure:

$$Y_i=[\mu_1(\boldsymbol{u}_i)+\varepsilon_i^{(1)}]\mathbb{I}(T_i=1)+[\mu_0(\boldsymbol{u}_i)+\varepsilon_i^{(0)}]\mathbb{I}(T_i=0),\quad 1\le i\le n,$$
$$\mu_1(\boldsymbol{u}_i)=2+f_{1,1}(u_{i,1})+f_{2,1}(u_{i,2})+f_{3,1}(u_{i,3})+f_{4,1}(u_{i,4})+f_{5,1}(u_{i,5})$$

and

$$\mu_0(\boldsymbol{u}_i)=1+f_{1,0}(u_{i,1})+f_{2,0}(u_{i,2})+f_{3,0}(u_{i,3})+f_{4,0}(u_{i,4})+f_{5,0}(u_{i,5}).$$

Similar to Huang et al. (2010), the covariate $\boldsymbol{u}_i=(u_{i,1},\ldots,u_{i,5})'$ is a 5×1 vector generated from a truncated multivariate normal with mean zero, covariance matrix $\boldsymbol{\Sigma}=(0.5^{|i-j|})_{5\times 5}$, and the truncated region is $[0,1]$. The setting of all $f_{j,t}(\cdot)$, $j=1,\ldots,5$, and $t\in\{0,1\}$ can be seen in Table 1. The treatment assignment T_i follows the Bernoulli distribution with probability $e(\boldsymbol{u}_i)=1/(1+e^{u_{i,1}-0.5u_{i,5}})$. The process of $\varepsilon_i^{(t)}$ is identically independent distributed and follows a standard normal distribution with a mean of zero and a variance of c_t, $t\in\{0,1\}$. The values of c_0 and c_1 are applied to control the signal-to-noise ratio, and we vary c_0 and c_1 such that $R^2=\mathrm{var}(\mu_1(\boldsymbol{u}_i))/[\mathrm{var}(\mu_1(\boldsymbol{u}_i))+\mathrm{var}(\varepsilon_i^{(1)})]=\mathrm{var}(\mu_0(\boldsymbol{u}_i))/[\mathrm{var}(\mu_0(\boldsymbol{u}_i))+\mathrm{var}(\varepsilon_i^{(0)})]$ ranges from 0.3 to 0.9.

Using these settings, we study the asymptotic optimality of our proposed procedure. We allow the candidate models in the treatment group and control group to be different. For the five covariates, we always include the first covariate but omit the last one to construct misspecified candidate models, which leads to the number of candidate models in treatment group and control group being both $2^3=8$. Thus,

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

the total number of candidate models is $K_n = 8 \times 8 = 64$. We consider the sample sizes for estimation as $n = 300$, 500, and 1,000.

Design 2: We consider the case where the number of covariates p diverges with the sample size. We generate data as follows:

$$Y_i(1) = 2 + \sum_{k=1}^{\lfloor 1.2n^{1/3} \rfloor} f_{k,1}(\boldsymbol{u}_i) + \varepsilon_i^{(1)},$$

$$Y_i(0) = 1 + \sum_{k=1}^{\lfloor 1.2n^{1/3} \rfloor} f_{k,0}(\boldsymbol{u}_i) + \varepsilon_i^{(0)}, \quad 1 \le i \le n,$$

where $\lfloor \cdot \rfloor$ denotes the floor function, and the subscript $\lfloor 1.2n^{1/3} \rfloor$ is the speed at which the dimension of $\boldsymbol{u}_i$ increases. We let $f_{k,t}(\boldsymbol{u}_i) = f_{k,t}(u_{i,k})$ for $k = 1, \ldots, 5$ and $t = 0, 1$ be the same as those in Design 1, and $f_{k,1}(\boldsymbol{u}_i) = -(k+1)^{1/2}(2 - u_{i,k} + 2u_{i,k-4})^2$ and

$$f_{k,0}(\boldsymbol{u}_i) = 5k^{1/2} \frac{\sin(2\pi(u_{i,k} + u_{i,k-3}))}{2 - \sin(2\pi u_{i,k})},$$

for $k = 6, \ldots, \lfloor 1.2n^{1/3} \rfloor$. We denote the number of covariates p as $\lfloor 1.2n^{1/3} \rfloor$. The first five covariates $u_1, \ldots, u_5$ are independently generated from $N(\mathbf{0}_{5\times 1}, \{0.5^{|i-j|}\}_{5\times 5})$ truncated to $[0, 1]$. The rest of covariates $u_6, \ldots, u_{\lfloor 1.2n^{1/3} \rfloor}$ are independently generated from $N(\mathbf{0}_{(p-5)\times 1}, \{0.5^{|i-j|}\}_{(p-5)\times(p-5)})$ truncated to $[-1, 0]$. The treatment assignment T_i follows the Bernoulli distribution with probability $e(\boldsymbol{u}_i) = 1/(1 + \exp(u_{i,1}^2 - 0.5u_{i,2} + u_{i,4}u_{i,6}))$, and $\{\varepsilon_i^{(1)}, \varepsilon_i^{(0)}\}$ are independent random errors generated from $\varepsilon_i^{(1)} \sim N(0, c_1)$ and $\varepsilon_i^{(0)} \sim N(0, c_0)$. Here, we vary c_0 and c_1 such that the values of R^2 are 0.3, 0.5, 0.7, and 0.9, where R^2 is defined in Design 1. The sample sizes are $n = 300$, 500, and 1,000. To reduce the computational burden, we employ the pre-screening method proposed by Fan et al. (2011) to order the covariates. After ordering, we construct the set of candidate models by including one additional covariate at each time based on the ordering. In this simulation study, sufficient dimension reduction is performed prior to matching, followed by nearest neighbor matching based on the low-dimensional nonlinear functions of $\boldsymbol{U}$. Specifically, generalized sliced inverse regression is employed to estimate the vectors in the central sufficient dimension reduction subspace, while cross-validated BIC (Li and Song, 2017) is utilized to determine the values of $r(1)$ and $r(0)$.

Design 3: In this setting, we mainly examine the estimation performance of our proposed method when the treatment effect is zero. We generate data from

$$Y_i(1) = \mu_1(\boldsymbol{u}_i) + \varepsilon_i^{(1)},$$

$$Y_i(0) = \mu_0(\boldsymbol{u}_i) + \varepsilon_i^{(0)}, \quad 1 \le i \le n,$$

where $\mu_1(\boldsymbol{u}_i) = 2 + \sum_{k=1}^{5} f_{k,1}(u_{i,k})$, and $f_{k,1}(\cdot)$ for $k = 1, 2.\ldots, 5$ are listed in Table 1. The covariates $\boldsymbol{u}_i$, for $i = 1, 2, \ldots, n$, are independently generated from

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

$N(\mathbf{0}_{5\times1}, \{0.5^{|i-j|}\}_{5\times5})$ truncated to $[0,1]$. The error $\varepsilon_i^{(1)} \sim N(0, c_1)$ and $\varepsilon_i^{(0)} \sim N(0, c_0)$, where c_1 and c_0 are set such that $R^2 = 0.7$. We let $\mu_0(\boldsymbol{u}_i) = \mu_1(\boldsymbol{u}_i) + A[\exp(0.6u_{i,1} - 0.75u_{i,2}) + 2\sin(2\pi u_{i,3}) + 2\log(u_{i,4}+1) + \sqrt{u_{i,5}+3}]$in this simulation study. When $A = 0$, it means that the treatment effect is zero for all individuals. The absolute value of the treatment effect increases when the absolute value of A increases.

In this design, we let $n = 1,000$, and $A = 0$, 0.5, 1, 2, and 3. We always include the first covariate but omit the last one in the five covariates to construct the candidate models. Furthermore, we add a constant treatment effect model to the candidate model set, which leads to the total number of candidate models being $K_n = 65$.

Design 4: We generate data as follows:

$$\begin{aligned} Y_i = {} & [\nu_1 + f_{1,1}(u_{i,1}) + f_{2,1}(u_{i,2}) + f_{3,1}(u_{i,3}) + f_{4,1}(u_{i,4}) + \varepsilon_i^{(1)}]\mathbb{I}(T_i = 1) \\ & + [\nu_0 + f_{1,0}(u_{i,1}) + f_{2,0}(u_{i,2}) + f_{3,0}(u_{i,3}) + f_{4,0}(u_{i,4}) + \varepsilon_i^{(0)}]\mathbb{I}(T_i = 0), \\ & 1 \le i \le n, \end{aligned}$$

where $\nu_1 = 2$, $\nu_0 = 1$, $f_{1,1}(u_1) = \sin(1.5u_1)$, $f_{2,1}(u_2) = u_2^3 + 1.5(u_2 - 0.5)^2$, $f_{3,1}(u_3) = -\phi(u_3, 0.5, 0.8^2)$, $f_{4,1}(u_4) = 2\sin(\exp(-0.5u_4))$, $f_{1,0}(u_1) = \cos(1.5u_1)$, $f_{2,0}(u_2) = -(u_2 - 1)^2$, $f_{3,0}(u_3) = (u_3 + 1)^2 + \cos(2\pi u_3)$, and $f_{4,0}(u_4) = 0$. The covariate vector $\boldsymbol{u}_i = (u_{i,1}, \ldots, u_{i,5})$ is a 5×1 vector generated from a truncated multivariate normal with mean 0.5, covariance matrix $\boldsymbol{\Sigma} = (0.5^{|i-j|})_{5\times5}$, and the truncated region is $[-1, 1]$. We consider randomized experiments and observational studies with propensity score functions of $e(\boldsymbol{u}_i) = 0.5$ and $e(\boldsymbol{u}_i) = 1/(1 + e^{u_{i,1}+u_{i,2}})$, respectively. The process of $\varepsilon_i^{(t)}$, for $t = 0, 1$, is the same as Design 1. In this setting, we consider the case where correct models (at least one) are included in the set of candidate models. We use all variables $u_1, \ldots, u_4$ to establish the candidate model set, and we include the first covariate in all candidate models. This procedure leads to $8 \times 8 = 64$ candidate models with two correct models in the candidate model set. The values of c_0 and c_1 are chosen to let $R^2 = 0.5$, 0.7, and 0.9. The sample sizes are $n = 500$, 1,000, 2,000, 3,000, and 4,000. To study the consistency of the weights, we calculate the sum of weights that are assigned to the correctly specified models.

5.2. Implementation and Methods to be Compared with

To implement our proposed procedure, we use five-fold CV and ten-fold CV to get the averaging weight vector. Similar to Rolling and Yang (2014), Rolling et al. (2019), and Zhao et al. (2024), the nearest neighbor matching with $M = 1$ is applied to estimate the unobserved potential outcome. We use the cubic B-spline with evenly spaced knots for all the nonlinear functions. The number of spline interior knots for each candidate model is chosen by five-fold CV. Cross-validated BIC (Li and Song, 2017) is applied to determine the reduction values $r(1)$ and $r(0)$. We consider the following estimators: (1) three model selection estimators (AIC, BIC,

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

and TECV); (2) three averaging estimators (SAIC, SBIC, and TEEM); (3) a causal forest estimator (labeled as "Forest"); (4) a double robust estimator (labeled as "DR"); and (5) a model averaging estimator by using J-fold CV criterion (labeled TECVMA). We find that the model averaging estimator with five-fold CV weights is almost the same as that with ten-fold CV through intensive simulations studies. Thus, it indicates that the finite sample performance of our proposed procedure is not sensitive to the choice of J. Due to space consideration, we only show the results of the model averaging estimator with five-fold CV weights.

We briefly discuss each estimator. The AIC and BIC scores of the kth candidate model are given by

$$\mathrm{AIC}_k = n\log(\hat{\sigma}_k^2) + 2\tilde{q}_k \quad \text{and} \quad \mathrm{BIC}_k = n\log(\hat{\sigma}_k^2) + \log(n)\tilde{q}_k,$$

where $\hat{\sigma}_k^2 = \|\hat{\boldsymbol{\Delta}} - \hat{\boldsymbol{\Delta}}^{(k)}\|^2/n$, and $\tilde{q}_k$ is the total number of basis functions in the kth candidate model. The TECV (Rolling and Yang, 2014) is a treatment effect CV aimed at minimizing treatment effect estimation errors. The AIC, BIC, and TECV model selection estimators are based on the model with the smallest AIC, BIC, and TECV scores, respectively. We define the SAIC and SBIC estimators as the model averaging methods, that, respectively, use the exponential AIC and BIC as the model weight. The weights of model k for the SAIC and SBIC estimators are constructed by

$$\omega_k = \frac{\exp(-\mathrm{AIC}_k/2)}{\sum_{s=1}^{K_n}\exp(-\mathrm{AIC}_s/2)} \quad \text{and} \quad \omega_k = \frac{\exp(-\mathrm{BIC}_k/2)}{\sum_{s=1}^{K_n}\exp(-\mathrm{BIC}_s/2)}.$$

The TEEM (Rolling et al., 2019) is a model combination method that uses weights of each candidate model based on its (estimated) accuracy for estimating treatment effect.

The causal forest estimator can be implemented by the R function "causal-forest" in the package "grf." Software implementations of causal forest have a number of parameters to allow users to fine-tune the algorithm's behavior, including the number of covariates to try at each split, and the number of trees. Similar to Qiu et al. (2020), the number of covariates to consider at every split is set as $\lfloor p/3 \rfloor$ for the forest algorithm, where p is the number of covariates and $\lfloor \cdot \rfloor$ denotes the floor function, and the number of trees in causal forest is set as 200.

The double robust estimator can be described as follows:

Step 1. Nuisance components estimation:

(a). construct estimates $\hat{e}(\boldsymbol{U})$ of the propensity scores $e(\boldsymbol{U})$ by logistic regression;

(b). construct estimates $(\hat{\mu}_1(\boldsymbol{U}), \hat{\mu}_0(\boldsymbol{U}))$ of the regression functions $(\mu_1(\boldsymbol{U}), \mu_0(\boldsymbol{U}))$ by the largest candidate model.

Step 2. Pseudo-outcome regression: Construct the pseudo-outcome

$$\Upsilon(\boldsymbol{U}) = \hat{\mu}_1(\boldsymbol{U}) - \hat{\mu}_0(\boldsymbol{U}) + \frac{T - \hat{e}(\boldsymbol{U})}{\hat{e}(\boldsymbol{U})(1 - \hat{e}(\boldsymbol{U}))}\{Y - \hat{\mu}_T(\boldsymbol{U})\}$$

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

and regress it on covariates $\boldsymbol{U}$ by spline regression, yielding

$$\hat{\Delta}_{\mathrm{dr}}(\boldsymbol{u}) = \hat{\mathrm{E}}\{\Upsilon(\boldsymbol{U})|\boldsymbol{U} = \boldsymbol{u}\}.$$

First, for Designs 1–3, we consider cases in which all candidate models are misspecified. For each simulation study, we compare the estimation performance of various methods by the average squared errors (ASEs). The ASEs are defined as

$$\mathrm{ASE}_j = \frac{1}{n_{\mathrm{new}}} \sum_{i=1}^{n_{\mathrm{new}}} (\hat{\Delta}(\boldsymbol{u}_i) - \Delta(\boldsymbol{u}_i))^2,$$

where $n_{\mathrm{new}} = 1,000$ is the size of the evaluation sample, $\hat{\Delta}(\boldsymbol{u}_i)$ is the estimator of $\Delta(\boldsymbol{u}_i)$ based on a training sample with n observations, and j denotes the jth replication. We compare the estimation performance of various methods using the normalized ASE (NASE), which is defined as

$$\mathrm{NASE} = \frac{1}{1000} \sum_{j=1}^{1000} \mathrm{ASE}_j / R_{0j},$$

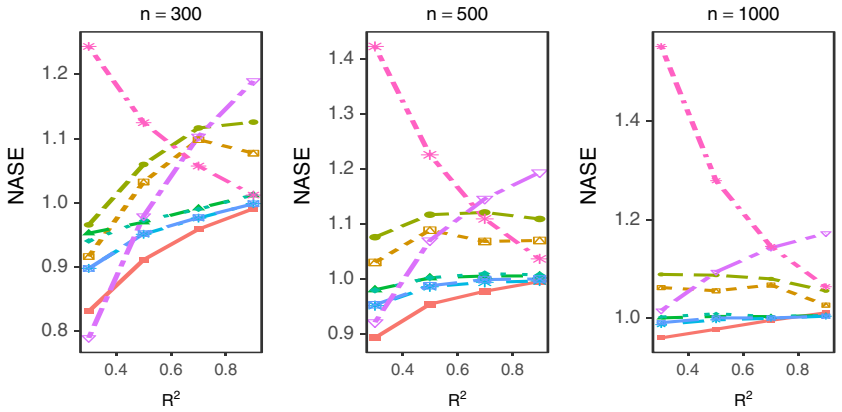
where R_{0j} is the ASE of the largest candidate model for the jth replication.

Second, we use Design 4 to verify the consistency of weight when there exists at least one correct model in the candidate model set. Here, we apply five-fold CV to choose the weight vector and then calculate the sum of weights that are assigned to the correctly specified models with different sample sizes and different values of R^2. The mean of $\hat{\Gamma}$ across 200 replications is shown in Figure 3.

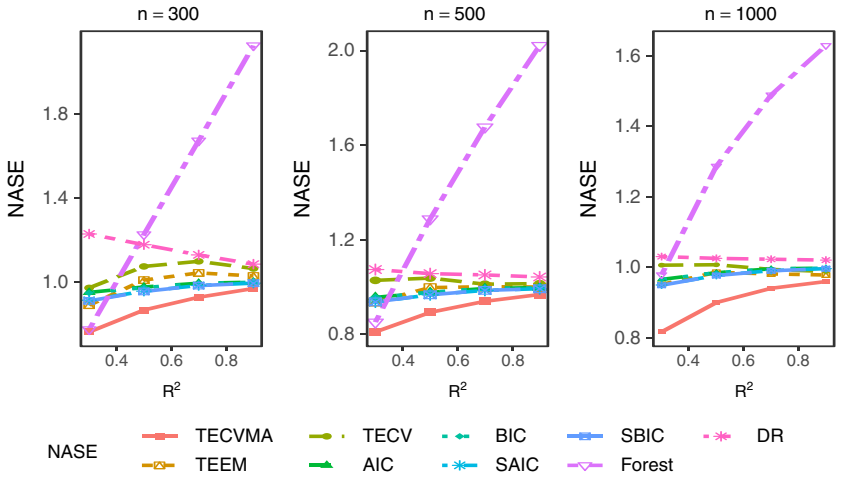
5.3. Simulation Results

We first examine the simulation results in Design 1 and Design 2, where all candidate models are misspecified. The results are shown in Figure 1. In each panel with different values of n, the NASE is displayed on the y-axis, and R^2 is displayed on the x-axis. From Figure 1, we note that in most cases, the TECVMA produces the smallest value of NASE among the competitors: TEEM, TECV, AIC, BIC, SAIC, SAIC, DR, and Forest. Notably, the efficacy of our method manifests superior performance, especially when the sample size is relatively large and the value of R^2 is low. The performance of causal forest is better than our method when R^2 and the sample size are relatively small, but it is worse in most other cases. Moreover, DR performs best when R^2 is large, but rather poorly when R^2 is relatively small. These empirical findings substantiate the overall superior behavior of our method in comparison to alternative approaches. With 1,000 simulation runs, the comparison results are reliable. For example, under Design 2 with $R^2 = 0.3$ and $n = 1,000$, the standard errors (SEs) of ASE_j/R_{0j} for TECVMA, TEEM, TECV, SAIC, SBIC, AIC, BIC, DR, and Forest are 0.002, 0.005, 0.005, 0.002, 0.002, 0.002, 0.002, 0.003, and 0.004, respectively. The simulation results demonstrate stability when

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press



(a) Design 1



(b) Design 2

FIGURE 1. NASE comparisons under Designs 1 and 2.

the number of simulation runs reaches 300, as evidenced by the quite small SE values.

Second, we evaluate the estimation performance of our proposed method when the treatment effect is zero or relatively small. These results are exhibited in Figure 2, where the NASE is displayed on the *y*-axis, and the absolute value of average treatment effect is displayed on the *x*-axis. From Figure 2, we find our proposed method performs worse than causal forest, TECV, and SBIC when the treatment effect is zero, but the difference between those methods is relatively small. As the absolute value of average treatment effect increases, our method

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

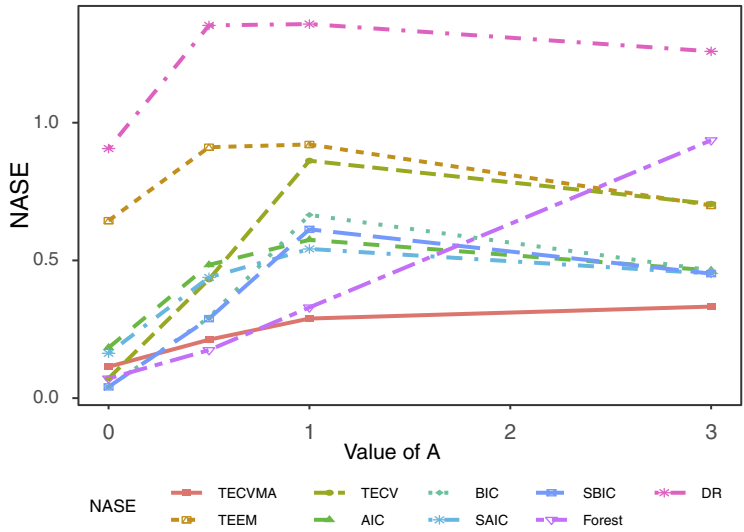
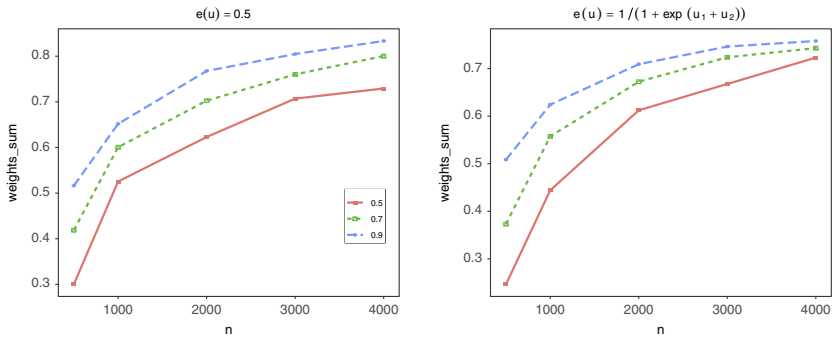


FIGURE 2. NASE comparisons under Design 3.



(a) Sum of weights assigned to correct models

(b) Sum of weights assigned to correct models

FIGURE 3. Assessing the convergence of weight vector $\hat{\boldsymbol{\omega}}$ under Design 4.

tends to produce the lower NASE than its rivals. In particular, when we enlarge the absolute value of treatment effect, the discrepancy between our method and the other methods becomes large. Overall, our approach seems to be most stable compared with others.

Finally, we use Design 4 to verify the consistency of weights in Theorem 2. The simulation results are shown in Figure 3, where we plot the sum of weights assigned to the correct models against the sample size under different values of R^2. We see that the sum of weights is monotonically increasing and converges to one as the sample size n increases. Moreover, we find that the convergence rate

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

of the sum of weights improves as R^2 becomes larger. These results support the conclusion of Theorem 2.

6. EMPIRICAL APPLICATION

In this section, we consider a data set from the European Central Bank's Consumer Expectation Survey (CES), which interviews about 10,000 households in the six largest euro area economies every month since April 2020. In September 2020, a special survey aimed at identifying the causal evidence that macroeconomic uncertainty affects overall household spending was implemented. In this survey, households were randomly assigned to one of five groups, in which certain types of information about macroeconomics were provided. Households in the control group received no information, while those in the other four groups were provided with four distinct types of information regarding macroeconomics. The detail of information provided for the third group (Treatment 2) is: "Professional forecasters are uncertain about economic growth in the euro area in 2021, with the difference between the most optimistic and the most pessimistic predictions being 4.8 percentage points. By historical standards, this is a big difference (Coibion et al., 2024)." To measure the causal effect of Treatment 2, we drop households that receive the other three treatment strategies. Each household provided the background information and some time-sensitive information on September and October 2020 surveys.

After deleting those observations with missing values, we obtain a sample containing $n = 2,081$ observations with a response variable (Y), a treatment index variable (T), and six covariate variables. We use the square root of total household spending in October 2020 as the outcome variable. The treatment index variable is $T \in \{0, 1\}$, where $T = 1$ for households who received the information about the amount of disagreement across professional forecasters, and $T = 0$ for households who received no information. The treatment group ($T = 1$) contains 276 households, with the remaining 1,805 households in the control group. Six covariates are used: the age of the households (U_1), employment situation (U_2), total combined net of tax income (U_3), household size (U_4), education (U_5), and the square root of the total household spending in past (U_6).

Model averaging estimator of CATE is constructed by assembling estimators from multiple additive candidate models. We allow treatment and control models to be different from each other. Moreover, we consider it mandatory to include the intercept, the total combined net of tax income, education, and past total household spending in both the treatment models and control models. Thus, the number of candidate models in the treatment group and control group are both $2^3 = 8$, which leads to the total number of additive candidate models being $K_n = 8 \times 8 = 64$. For each candidate model, terms involving categorical variables (employment status, education, and household size) are linear and the cubic B-spline is applied with evenly distributed knots for all the nonlinear

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

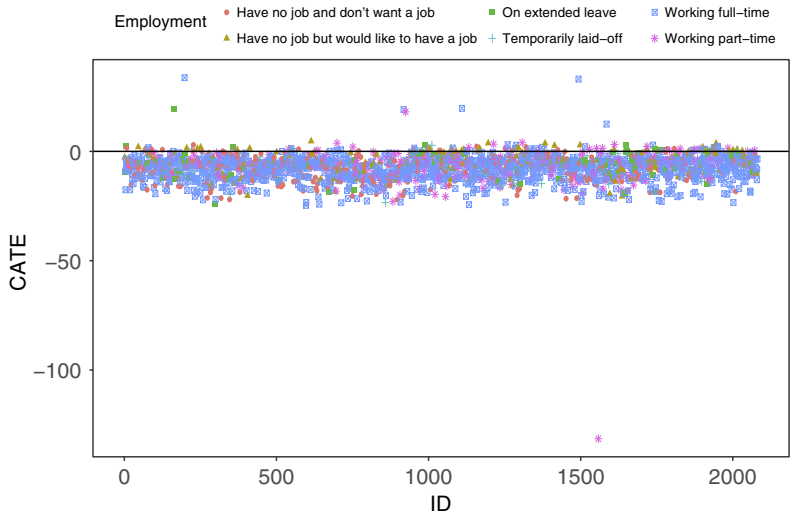


FIGURE 4. Model averaging estimates for heterogeneous treatment effect.

functions. The number of spline interior knots for each candidate model is chosen by five-fold CV. We also use five-fold CV to choose the model averaging weights.

The model averaging estimates of heterogeneous treatment effects are graphically represented in Figure 4 on the y-axis against the individual index. Clearly, the estimates of CATE predominantly fall below zero, indicating that the informed macroeconomic uncertainty would lead to a significant reduction in household's total monthly expenditures in subsequent months. This is consistent with the result of Coibion et al. (2024). We further calculate the model averaging estimate of CATE over the range of income and the total household spending in the past for this data. Here, we set the employment status at working full-time, household size at 2, and age at 40. The corresponding results are shown in Figure 5 (a) and (b) for primary and tertiary, respectively. We see that the informed macroeconomic uncertainty had more impact on those with large incomes and low total household spending in the past.

Given the challenge that the difference between $Y_i(1)$ and $Y_i(0)$ is inherently unobservable, we employ a "data guided simulation" (DGS) approach to evaluate the performance of various methods in estimating CATE under different but sensible simulation scenarios. This methodology follows the experimental framework established in Rolling and Yang (2014), Rolling et al. (2019), and Nie and Wager (2021). Each simulation scenario is constructed based on the model selected through one of the model selection methods on the original data. For each simulation scenario, we generate a "true" potential outcome, $(Y_i^*(1), Y_i^*(0))$, utilizing the six covariates values and the selected

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

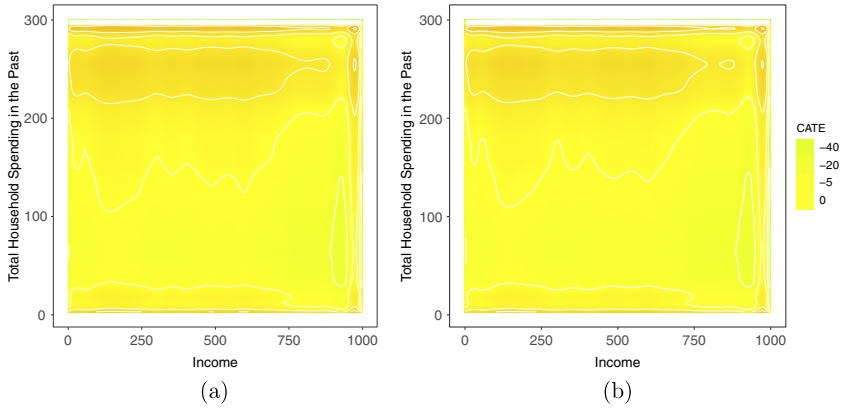


FIGURE 5. Contour plots for $\hat{\Delta}(\boldsymbol{u})$ for CES data.

model. Specifically, while both the covariates and treatment selection are available from the original data, we generate $(Y_i^*(1), Y_i^*(0))$ by incorporating independent and identically distributed noise into the estimated regression functions obtained from the selected model. This noise is drawn from a mean-zero normal distribution, with its variance derived from the selected model.

After the new data are generated, we apply three model selection methods and four combination approaches, along with causal random forest and double robust estimators, to $(Y_i^*(1), Y_i^*(0), T_i, \boldsymbol{U}_i)$, $1 \leq i \leq 2,081$, to obtain an estimate of $\Delta^*(\boldsymbol{u})$. Given that the true value of $\Delta^*(\boldsymbol{u})$ is known in each simulation scenario, we can systematically evaluate and compare the performances of the estimators derived from model selection, model averaging, causal random forest, and double robust methods, based on the respective truth versions generated by AIC, BIC, and TECV model selection approaches. For each simulation scenario, we omit the last variable to mimic the situation where all candidate models are misspecified. Subsequently, we assess the risk associated with each $\hat{\Delta}^*(\boldsymbol{u}_i)$ under the squared error loss function by computing the ASE across the $n = 2,081$ sample values of $\boldsymbol{u}_i$.

The mean and median of ASE, and the SE of $\hat{\Delta}^*(\boldsymbol{u}_i)$ across 400 distinct realizations are presented in Table 2. Our proposed methodology exhibits superior performance compared to the alternative approaches, particularly in terms of both the mean and median values of ASE. Notably, when $Y_i^*(1)$ and $Y_i^*(0)$ are determined through the TECV model selection process, the mean of ASE of TECVMA demonstrates approximately 15% reduction compared to that of TECV. These empirical findings substantiate that the proposed procedure constitutes an effective instrument for estimating heterogeneous treatment effects.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Table 2. DGS results: Estimated risk of $\Delta^*(\boldsymbol{u})$ under three scenarios consistent with CES data

Method	ASE	Model selection method determining $(Y_i^*(1), Y_i^*(0))$		
		AIC	BIC	TECV
TECVMA	Mean	**41.73**	**41.73**	**43.22**
	Median	**40.94**	**40.94**	**42.22**
	SE	**0.15**	**0.15**	**0.19**
TEEM	Mean	44.13	44.13	46.98
	Median	42.22	42.22	43.69
	SE	0.42	0.42	0.55
TECV	Mean	46.78	46.78	49.60
	Median	45.15	45.15	47.21
	SE	0.38	0.38	0.49
SAIC	Mean	45.36	45.36	48.25
	Median	44.39	44.39	46.04
	SE	0.36	0.36	0.46
SBIC	Mean	45.18	45.18	48.09
	Median	44.32	44.32	46.17
	SE	0.29	0.29	0.39
AIC	Mean	46.92	46.92	50.02
	Median	45.39	45.39	47.52
	SE	0.37	0.37	0.46
BIC	Mean	46.26	46.26	49.10
	Median	45.15	45.15	47.05
	SE	0.28	0.28	0.37
Forest	Mean	84.64	84.64	93.03
	Median	78.51	78.51	86.29
	SE	1.32	1.32	1.66
DR	Mean	47.11	47.11	51.91
	Median	45.47	45.47	49.22
	SE	0.45	0.45	0.79

7. CONCLUSION

This article offers a model averaging method to achieve an accurate and stable estimation for heterogeneous treatment effects using additive causal models. We propose J-fold CV to select the data-driven weights, and obtain an asymptotic optimality or consistency of both weighting and the model averaging estimator based on additive spline models. Numerical results strongly support that our model averaging approach is often better than model selection and the other model averaging methods.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Our proposed procedure is theoretically justified under two assumptions: unconfounded assignment and overlap. Although these are common conditions in most literature on causal inference, they may be violated in practice. For instance, the number of observed individuals in the treatment group may sometimes be very small in observational studies. Future work will be focused on handling such imbalanced data. In addition, deriving asymptotic distributions of the model averaging estimators remains an important but challenging research task.

Supplementary Material

Na Li, Yu Fei, Yuhong Yang, and Xinyu Zhang (2025): Supplement to "Averaging Estimators of Heterogeneous Treatment Effects under Additive Models," Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466625100200.

REFERENCES

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

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

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113, 7353–7360.

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics*, 17, 453–510.

Cai, T., Tian, L., Wong, P. H., & Wei, L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12, 270–278.

Claeskens, G., Croux, C., & Van Kerckhoven. (2006). Variable selection for logistic regression using a prediction-focused information criterion. *Biometrics*, 62, 972–979.

Coibion, O., Georgarakos, D., Gorodnichenko, Y., Kenny, G., & Weber, M. (2024). The effect of macroeconomic uncertainty on household spending. *American Economic Review*, 114, 645–677.

Cui, G., Li, N., Wan, A., & Zhang, X. (2025). Model averaging for estimating treatment effects with binary responses. *Applied Stochastic Models in Business and Industry*, 41, 2898–2916.

Fan, J., Feng, Y., & Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106, 544–557.

Fang, F., Li, J., & Xia, X. (2022). Semiparametric model averaging prediction for dichotomous response. *Journal of Econometrics*, 229, 219–245.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75, 1175–1189.

Hansen, B. E., & Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167, 38–46.

Hoeting, J., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.

Hu, X., & Zhang, X. (2023). Optimal parameter-transfer learning by semiparametric model averaging. *Journal of Machine Learning Research*, 24, 1–53.

Huang, J., Horowitz, J. L., & Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38, 2282–2313.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 7, 443–470.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Koch, B., Vock, D. M., & Wolfson, J. (2018). Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, 74, 8–17.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116, 4156–4165.

Lee, K. Y., Li, B., & Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Annals of Statistics*, 41, 221–249.

Lewbel, A., Choi, J. Y., & Zhou, Z. (2023). Over-identified doubly robust identification and estimation. *Journal of Econometrics*, 235, 25–42.

Li, B., Artemiou, A., & Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Annals of Statistics*, 39, 3182–3210.

Li, B., & Song, J. (2017). Nonlinear sufficient dimension reduction for functional data. *Annals of Statistics*, 45, 1059–1095.

Li, J., Lv, J., Wan, A. T. K., & Liao, J. (2020). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association*, 117, 495–509.

Li, J., Yu, T., Lv, J., & Lee, M. (2021). Semiparametric model averaging prediction for lifetime data via hazards regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70, 1187–1209.

Liao, J., Wan, A. T., He, S., & Zou, G. (2023). Frequentist model averaging for the nonparametric additive model. *Statistica Sinica*, 33, 401–430.

Liao, J., Zong, X., Zhang, X., & Zou, G. (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics*, 209, 35–60.

Liu, C.-A. (2015). Distribution theory of the least squares averaging estimator. *Journal of Econometrics*, 186, 142–159.

Liu, C.-A., & Kuo, B.-S. (2016). Model averaging in predictive regressions. *The Econometrics Journal*, 19, 203–231.

Luo, W., & Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika*, 103, 875–887.

Luo, W., & Zhu, Y. (2020). Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, 38, 888–900.

Lv, J., & Liu, J. S. (2014). Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 141–167.

Min, L., Sadiq, S., Feaster, D. J., & Ishwaran, H. (2018). Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics*, 27, 209–219.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108, 299–319.

Opsomer, J. D., & Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25, 186–211.

Qiu, Y., Xie, T., Yu, J., & Zhang, X. (2020). Mallows-type averaging machine learning techniques. Working Paper.

Racine, J. S., Li, Q., Yu, D., & Zheng, L. (2023). Optimal model averaging of mixed-data kernel-weighted spline regressions. *Journal of Business & Economic Statistics*, 41, 1251–1261.

Ravikumar, P., Lafferty, J., Liu, H., & Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71, 1009–1030.

Rolling, C. A., & Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 749–769.

Rolling, C. A., Yang, Y., & Velez, D. (2019). Combining estimates of conditional treatment effects. *Econometric Theory*, 35, 1089–1110.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, 1346–1370.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press

Schumaker, L. (1981). *Spline functions: Basic theory*. Wiley.
Splawa-Neyman, J., Dabrowska, D. M., & Speed, T. (1990). On the application of probability theory to agricultural experiments. *Statistical Science*, 5, 465–472.
Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13, 689–705.
Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113, 1228–1242.
Xue, L. (2009). Consistent variable selection in additive models. *Statistica Sinica*, 19, 1281–1296.
Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, 74, 135–161.
Yuan, C., Wu, Y., & Fang, F. (2022). Model averaging for generalized linear models in fragmentary data prediction. *Statistical Theory and Related Fields*, 6, 344–352.
Zhang, X., Yu, D., Zou, G., & Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111, 1775–1790.
Zhang, X., Zou, G., & Liang, H. (2014). Model averaging and weight choice in linear mixed-effects models. *Biometrika*, 101, 205–218.
Zhao, Z., Zhang, X., Zou, G., Wan, A. T., & Tsoc, G. K. (2024). Model averaging for estimating treatment effects. *Annals of the Institute of Statistical Mathematics*, 76, 73–92.
Zou, J., Wang, W., Zhang, X., & Zou, G. (2022). Optimal model averaging for divergent-dimensional poisson regressions. *Econometric Reviews*, 41, 775–805.

https://doi.org/10.1017/S0266466625100200 Published online by Cambridge University Press